Des apprentis pour le traitement automatique des langues

MÉMOIRE

Mémoire présenté pour l'obtention de l'

Habilitation à Diriger des Recherches

de l'Université Pierre et Marie Curie - Paris VI

(Spécialité Informatique)

par

François Yvon

Composition du Jury

Président

Patrick Gallinari, Professeur à l'Université de Paris VI

Rapporteurs

Walter Daelemans, Professeur à l'Université d'Anvers (Belgique)
Pierre Dumouchel, Professeur à l'École de Technologie Supérieure de l'Université du Québec (Canada)
Marc El-Bèze, Professeur à l'Université d'Avignon

Examinateurs

Rémi Gilleron, Professeur à l'Université de Lille III
Benoît Habert, Professeur à l'Université de Paris X
Ludovic Lebart, Directeur de recherches CNRS
Joseph Mariani, Directeur de recherches CNRS

Remerciements

Je remercie sincèrement les membres du jury pour leurs commentaires sur ce mémoire, qui m'auront permis de l'améliorer significativement, et en particulier :

Monsieur Patrick Gallinari, Professeur à l'Université Paris VI, pour m'avoir accompagné tout au long de ce travail et pour m'avoir fait l'honneur de présider mon jury.

Monsieur Walter Daelemans, Professeur à l'Université d'Anvers, Monsieur Marc Elbèze, Professeur à l'Université d'Avignon, et Monsieur Pierre Dumouchel, Professeur à l'Ecole de Technologie Supérieure à l'Université du Québec de m'avoir fait l'honneur d'être rapporteurs de mon mémoire d'habilitation.

Monsieur Benoît Habert, Professeur à l'Université de Paris X, pour avoir encouragé et critiqué mon travail sur l'analogie et pour avoir accepté d'être membre de mon jury.

Monsieur Rémi Gilleron, Professeur à l'Université de Lille III et Monsieur Joseph Mariani, Directeur de Recherches au CNRS, d'avoir accepté d'être membre de mon jury.

Monsieur Ludovic Lebart, Directeur de Recherches au CNRS, que je côtoie depuis le début de ma carrière de chercheur : sa participation à mon jury constitue un très grand honneur.

2

Table des matières

1 TAL et Apprentissage Automatique **11**

1.1 Introduction . . . 12

1.1.1 Le traitement automatique des langues . . . 12

1.1.2 L'apprentissage automatique . . . 15

1.1.3 Une convergence inévitable ? . . . 19

1.1.4 Conclusion . . . 22

1.2 Une typologie des applications de l'apprentissage en traitement automatique des langues . . . 22

1.2.1 Les problèmes d'apprentissage non-supervisé . . . 23

1.2.2 Les problèmes de catégorisation . . . 24

1.2.3 Les tâches « séquentielles » . . . 25

1.2.4 Problèmes dérivés . . . 26

1.2.5 Conclusion . . . 28

1.3 Des questions difficiles . . . 28

1.3.1 Des données . . . 28

1.3.2 Le choix d'une représentation . . . 32

1.3.3 Estimation des modèles . . . 34

1.3.4 Quelques perspectives . . . 35

1.4 Un bilan contrasté . . . 37

1.4.1 Des évolutions significatives . . . 38

1.4.2 Un mariage de convenance ? . . . 41

2 Analogies **51**

2.1 De l'analogie . . . 52

2.1.1 L'analogie en intelligence artificielle et en sciences cognitives . . . 52

2.1.2 L'analogie en linguistique . . . 57

2.1.3 L'analogie en traitement automatique des langues . . . 60

2.1.4 Pour une approche analogique . . . 62
2.2 Un modèle d'inférence analogique . . . 65
2.2.1 Rapports de proportion . . . 65
2.2.2 Transfert analogique . . . 66
2.2.3 Recherche et comparaison de motifs analogiques . . . 67
2.3 Quelques expérimentations . . . 68
2.3.1 Prononciation de mots inconnus . . . 68
2.3.2 Analyse morphologique de mots inconnus . . . 70
2.3.3 Discussion . . . 73
2.3.4 Conclusion . . . 75
2.4 Analogies généralisées . . . 76
2.4.1 Une nouvelle définition de l'analogie . . . 76
2.4.2 Un solveur à états finis . . . 77
2.4.3 Compléments et généralisations . . . 79
2.5 Conclusion et perspectives . . . 83
2.5.1 Travail expérimental . . . 84
2.5.2 Vers des modèles graduels de l'analogie ? . . . 84

3 Exploitation statistique de corpus 93
3.1 Acquisition de paraphrases . . . 94
3.1.1 Cadre général . . . 94
3.1.2 Apprentissage de relations de paraphrase . . . 95
3.2 Catégorisation automatique de documents . . . 97
3.2.1 Cadre général . . . 98
3.2.2 Classifier pour catégoriser . . . 99
3.2.3 Conclusion . . . 100
3.3 Classification de documents . . . 101
3.3.1 Le modèle . . . 101
3.3.2 Quelques observations empiriques . . . 102
3.3.3 Améliorer l'estimation . . . 102
3.3.4 Un bilan et quelques perspectives . . . 103
3.4 Conclusion . . . 104

4 Contributions au traitement de la parole 109

4.1 Translittération orthographique-phonétique . . . 109
4.1.1 Les systèmes à base de règles . . . 110
4.1.2 Conclusion et perspectives . . . 114
4.2 Reconnaissance de la parole . . . 115
4.2.1 Introduction : un rapide point sur l'état de l'art . . . 115
4.2.2 Reconnaissance de noms propres . . . 118
4.2.3 Autour de Sirocco . . . 119
4.2.4 Une stratégie d'optimisation du graphe de décodage . . . 122
4.2.5 Perspectives : vers une meilleure utilisation des ressources . . . 123

A Quelques tâches d'analyse linguistique **133**
A.0.6 Segmentation et prétraitements . . . 133
A.0.7 Analyse lexicale, ressources lexicales . . . 134
A.0.8 Analyse syntaxique : petit à petit... . . . 137
A.0.9 Sémantique . . . 140

B Sélection d'articles **147**

Avant Propos

> Mais on peut bien dire que c'est un vrai fainéant que cet Antoine, et son « Antoinesse » ne vaut pas mieux que lui, ajoutait Françoise, qui, pour trouver au nom d'Antoine un féminin qui désignât la femme du maître d'hôtel, avait sans doute dans sa création grammaticale un inconscient ressouvenir de chamoine et de chamoinesse.
>
> M. Proust, *le Côté de Guermantes* (p. 27 dans l'édition Folio)

Contexte

Ce document présente une synthèse de mes activités de recherche dans le domaine des technologies de la langue. Statisticien d'origine, formé à l'école de l'analyse des données auprès d'A. Morineau, de T. Lambert et de L. Lebart, j'ai été introduit à ce domaine par A. Bonnet, grâce auquel j'ai pu faire un doctorat à l'ENST, puis, remplaçant au pied levé M. Rajman, j'ai pu ensuite intégrer le corps enseignant de cette école.

Une telle synthèse n'est pas aisée : au cours de ces quelques dix ans, comme enseignant et comme chercheur, j'ai été amené à aborder ces technologies sous de multiples aspects et à devoir plus d'une fois emjamber les frontières symboliques qui séparent et structurent les différentes (sous-)communautés de recherche travaillant sur ces technologies :

– frontière entre les méthodes « à base de règles », que j'enseigne, et les méthodes « à base de données », sur lesquels portent presqu'exclusivement mes travaux de recherche ;
– frontière entre les méthodes d'apprentissage statistique et les méthodes symboliques, qu'une partie de mon travail de recherche vise à concilier ;
– frontière entre analyse linguistique d'une part et analyse de données linguistiques, qui recouvrent des réalités différentes (mais l'une peut-elle se passer de l'autre ?) ;
– frontière entre traitement de la langue écrite et traitement de la langue orale, deux domaines auxquels je suis également attaché et pour la « réconciliation » desquels je m'efforce de militer.

Cette synthèse aurait été impossible si, au fil de ses années, on n'avait assisté à un rapprochement continu entre les différents domaines : convergence des méthodes de traitement de l'écrit et de l'oral ; envahissement progressif de l'ensemble du spectre du traitement automatique des langues par des méthodes « à base de données » initialement confinées aux domaines du traitement de la parole et de la recherche d'information. C'est, je le ne cache pas, avec un soulagement certain que j'ai assisté (beaucoup) et contribué (un peu) à ces évolutions, qui m'ont permis de progressivement ramener ma schizophrénie professionnelle à un niveau aujourd'hui presqu'acceptable.

Si donc, *a posteriori*, cette synthèse est devenue réalisable (chapitre 1), son organisation n'a que pu conserver la trace de ces aller-retours entre différentes problématiques, les unes plus applicatives, les autres plus théoriques ; les unes plus inspirées de mes travaux sur l'écrit (chapitres 3 et 2), les autres de mes travaux sur l'oral (chapitre 4). Le guide de lecture détaillé qui suit permettra, je l'espère, à chacun, selon ses préoccupations,

de tracer son chemin dans ce document.

Guide de lecture

Ce document est organisé en quatre chapitres principaux, qui sont complétés par plusieurs annexes.

Le premier chapitre, entièrement original, est consacré à une présentation de l'état de l'art en matière d'utilisation de méthodes d'apprentissage automatique pour accomplir des tâches de traitement automatique des langues. Cette présentation se focalise principalement sur les méthodes statistiques, qui constituent aujourd'hui le courant dominant en apprentissage automatique. Après une rapide présentation de ces deux domaines (section 1.1) et une analyse des circonstances qui ont favorisé leur convergence, je reviens, à la section 1.2, sur un certain nombre de tentatives visant à appliquer des méthodes statistiques à des données linguistiques. Je considère successivement des tâches d'apprentissage non-supervisé et supervisé, en distinguant le cas particulier des données structurées, en particulier séquentielles, dont le traitement est une des principales spécificités du traitement automatique des langues. Au terme de ce tour d'horizon, je tente, à la section 1.3, de dégager une synthèse raisonnée des principales difficultés que ces tentatives rencontrent, ainsi que des méthodes et outils qui ont progressivement été élaborées pour y faire face. Ce chapitre se conclut (section 1.4) par un bilan somme toute assez mitigé de ces travaux.

Les trois chapitres qui suivent sont consacrés à une synthèse de mes propres travaux de recherche, ventilés sur trois axes principaux : les contributions à l'apprentissage par analogie (chapitre 2), les contributions à l'apprentissage statistique (chapitre 3) et les contributions au traitement automatique de la parole (chapitre 4). Ces trois chapitres sont complètement indépendants les uns des autres ; ils s'achèvent chacun par un bilan du travail accompli et par une présentation de quelques perspectives.

Ces chapitres sont de longueurs inégales : les travaux portant sur l'apprentissage par analogie, qui font l'objet du chapitre 2, constituent la partie la plus personnelle (et, partant, la plus longue) de ce document. Prenant acte de la marginalité relative de ces travaux, ce chapitre débute par une longue justification de l'intérêt de ces méthodes, s'appuyant sur une description de l'état de l'art en la matière (section 2.1). Je présente ensuite un cadre général pour l'apprentissage par analogie (section 2.2), originellement développé durant mon travail de thèse et consolidé par une collaboration soutenue avec V. Pirrelli. Une première implantation de ce modèle a donné lieux à diverses expérimentations pour des tâches d'analyse lexicales, qui sont présentées et discutées à la section 2.3. Je discute enfin une implantation plus récente de ce modèle, qui s'appuie, en particulier, sur une définition générale des relations analogiques élaborée dans le cadre du travail de thèse de N. Stroppa et sur une implantation du calcul de relations analogiques utilisant des transducteurs finis (section 2.4). J'évoque, pour conclure, quelques résultats expérimentaux obtenus dans ce cadre ainsi que les nombreuses perspectives sur lesquelles ce travail débouche.

Les deux chapitres qui suivent sont, en comparaison, beaucoup plus succints. La rapide description de mes contributions à l'apprentissage statistique, qui fait l'objet du chapitre 3, s'organise autour de la présentation de trois études portant sur des problèmes de gestion documentaire (au sens large), réalisées dans le cadre de collaborations industrielles. Ces étude ont été menées en étroite collaboration avec des doctorants de l'ENST et, le cas échéant, leur(s) co-encadrant(s). Sont ainsi successivement présentés un travail portant sur l'acquisition automatique de paraphrases (thèse de F. Duclaye, à la section 3.1) ; un travail portant sur des applications de la catégorisation de documents dans un contexte de traitement semi-automatique de courriers électroniques (thèse de R. Vinot, section 3.2) ; enfin un travail portant sur la classification non-supervisée de collections de documents (thèse de L. Rigouste, section 3.3).

Le chapitre 4 présente un résumé de mes contributions au traitement automatique de la parole, en distinguant les travaux portant sur la transcription orthographique-phonétique pour la synthèse vocale (développement de systèmes de transcription et de dictionnaires), qui font l'objet de la section 4.1 ; et les travaux portant sur

les systèmes de reconnaissance vocale (développement de décodeurs de parole, mise au point de systèmes de reconnaissance) à la section 4.2.

Ce mémoire est complété par deux annexes : la première est un complément, à destination des lecteurs peu familiers avec le traitement automatique des langues, du chapitre 1. ELle est consacrée à une présentation de quelques grandes tâches du de l'analyse linguistique et des façons traditionnelles de les envisager. La seconde annexe (non reproduite dans le document de synthèse final) est une présentation factuelle de l'ensemble de mes activités d'enseignement, d'encadrement et de recherche.

Remerciements (bis)

Le succès d'un travail de recherche est pour beaucoup une question de rencontres : rencontres de problèmes, rencontres de personnes : j'ai eu la chance d'en faire de magnifiques, que ce soient des étudiants (élèves, stagiaires, doctorants) souvent brillants et stimulants que j'ai eu la chance d'encadrer ; que ce soit avec des collègues dans les différentes communautés (traitement automatique des langues, traitement de la parole, analyse de données textuelles, apprentissage automatique) que j'ai, de près ou de loin, fréquentées. Un grand merci d'ensemble à tous. J'essaierai, dans les chapitres qui suivent, de rendre leur dû à ceux avec lesquels les collaborations ont été les plus fortes.

La poursuite d'une activité de recherche, surtout quand elle aborde des domaines parfois marginaux (dans une institution d'enseignement supérieur technique) tels que ceux qui sont évoqués dans ce document, est d'autant plus aisée que l'institution y est favorable : c'est encore le cas de l'ENST, grâces en soient rendues en particulier à B. Robinet, « doyen » des enseignants-chercheurs de l'ENST et à J.-P. Tubach et H. Maître, directeurs successifs du Laboratoire CNRS "Traitement et Communication de l'Information", qui ont soutenu avec bienveillance l'activité de l'école en traitement automatique des langues ; ainsi qu'à U. Finger et à M. Riguidel, qui, se succédant à la tête du département "Informatique" (puis "Informatique et Réseaux") de l'ENST, ont contribué à maintenir vivant cet axe de recherche.

Ces années passées à l'ENST m'ont permis de mesurer la chance que j'avais d'évoluer au quoditien dans un environnement humain chaleureux et amical au sein des équipes de recherche auxquels j'appartiens, qu'en soient remerciés J.-L. Dessalles, A. Grumbach, E. Lecolinet, J.-M. Saglio, T. Abdessalem et G. Hébrail, mes collègues du groupe "Information, Communication, Complexité" ; ainsi qu'O. Cappé et E. Moulines qui m'accueillent (à temps partiel) au sein du groupe "Traitements Statistiques et Applications aux Communications" et, last but not least, G. Richard et G. Chollet, avec qui je partage de nombreuses activités en traitement de la parole.

La recherche n'occupant qu'une partie de mon temps, qu'il me soit également permis de dire ma gratitude à mes autres collègues du département "Informatique et Réseaux" (et d'ailleurs), avec lesquels je partage de nombreuses activités d'enseignement, en particulier O. Hudry, I. Charon, J. Sakarovitch, A. Demaille, I. Demeure et L. Pautet. Merci également aux équipes administrative (S. Béranger, H. Soussou, C. Bizart) et technique (P. Dax, S. Gadret, H. Aissaoui, G. Mouret, P. Busch) du département "Informatique et Réseaux" pour leur disponibilité et leur soutien discret et efficace.

Notations

V	un ensemble
$\|V\|$	le cardinal de l'ensemble V
X	une variable aléatoire (le plus souvent discrète)
$\mathcal{X}$	l'ensemble des réalisations possibles de X
x	une réalisation de X
$P(X = x), P(x)$	la probabilité de la réalisation x de X
$P(X = x\|Y = y), P(x\|y)$	la probabilité conditionnelle que $X = x$ sachant $Y = y$
$\mathbf{X}$	une variable aléatoire m-dimensionnelle
$\mathbf{x}$	un vecteur m-dimensionel d'observations
$\|\mathbf{x}\|$	la norme L^2 du vecteur $\mathbf{x}$: $\|\mathbf{x}\| = \sum_i x_i^2$
X	une variable aléatoire séquentielle
x	une séquence d'observations $x = x_1 \dots x_t \dots x_l$
x	une séquence d'observations m-dimensionnelles
$\mathcal{C}$	un ensemble i.i.d d'observations $\mathcal{C} = \{x^{(1)} \dots x^{(n)}\}$
$\updownarrow(\mathcal{C}\|\theta)$	la vraisemblance d'un corpus pour le modèle paramétré par θ
$\mathcal{L}(\mathcal{C}\|\theta)$	la log-vraisemblance : $\mathcal{L}(\mathcal{C}\|\theta) = \log(\updownarrow(\mathcal{C}\|\theta))$
$\mathrm{N}(e)$	la fréquence de la réalisation e de la variable E dans un corpus
Σ	un alphabet, inventaire fini de symboles
$\Sigma^\star$	le langage universel sur l'alphabet Σ
ε	le mot vide
w	un mot sur un alphabet Σ
$\|w\|$	la longueur d'un mot de $\Sigma^\star$ ($\|\varepsilon\| = 0$)
w_1^n	un mot sur un alphabet Σ, vu comme une suite $w_1 \dots w_n$ de symboles de Σ
$w(i)$	le ième symbole de w, pour $i \in [1 \dots \|w\|]$
vw	la concaténation des mots v et w
I	un ensemble ordonné d'entiers (d'indices)
$w(I)$	le sous-mot de w construit en assemblant les symboles de w aux positions I, pour $I \subset \{1 \dots \|x\|\}$
$v \in w$	v est un sous-mot de w

Chapitre 1

Traitement des Langues et Apprentissage Automatique : un tour d'horizon

Une analyse rétrospective des évolutions récentes du traitement automatique des langues fait apparaître deux tendances marquées : une forte diffusion de méthodologies et de techniques empruntées au domaine de l'apprentissage automatique, d'une part ; un rapprochement des thématiques du traitement automatique des langues avec celles du traitement automatique de la parole (TAP) et de la recherche d'information (RI)[1]. Les exemples de ces rapprochements sont multiples : pour n'en prendre qu'un seul, la recherche en traduction automatique, application emblématique du traitement automatique des langues, suscite un intérêt croissant en reconnaissance vocale (traduction parole-parole) comme en recherche d'information multilingue, conduisant, dans une certaine mesure, à une reformulation des problèmes et à de nouvelles propositions pour les résoudre.

En guise de d'introduction générale au cadre dans lequel se sont déroulées mes recherches, ce chapitre a l'ambition de donner une vue d'ensemble des applications des méthodes d'apprentissage en traitement automatique des langues et dans les domaines apparentés[2]. Compte-tenu de l'étendue du domaine et du foisonnement des travaux, cette ambition peut paraître démesurée. Pourtant, si les ambitions théoriques et les cadres applicatifs qui motivent le recours aux méthodes de l'apprentissage sont très divers, si les types de données et de langues qui sont traitées par ces méthodes sont très variés, la plupart des travaux repose sur l'utilisation d'un nombre relativement réduit de méthodes et modèles. À partir d'une cartographie de ces méthodes, j'essaierai de montrer la communauté des difficultés qui se posent à toutes les applications des méthodes d'apprentissage en traitement automatique des langues et de discuter de la manière dont ces difficultés sont abordées, traitées et résolues.

Cette discussion est organisée comme suit : une section introductive présente rapidement les domaines du traitement automatique des langues et de l'apprentissage automatique, en particulier des méthodes probabilistes. Cette introduction fournira l'occasion de préciser certains termes et de fournir quelques concepts de base. La lecture de cette introduction sera, le cas échéant, utilement complétée par l'annexe A, qui décrit un certain nombre de tâches du traitement automatique des langues, ainsi que par les encadrés grisés qui détaillent le point de vue probabiliste sur certains problèmes d'apprentissage.

La section 1.2 dresse un large panorama des applications de l'apprentissage automatique en traitement

[1] Il n'est d'ailleurs pas interdit de voir dans la première de ces tendances une conséquence plus ou moins directe de la seconde, dans la mesure où les méthodes statistiques sont historiquement des outils de base du traitement de la parole et de la recherche d'information.

[2] Pour l'essentiel, cette présentation de l'état de l'art restera centrée sur le domaine du traitement automatique des langues et ne fera qu'occasionnellement allusion aux travaux qui sont plus au centre des préoccupations du TAP (le traitement et la modélisation du signal de parole) ou de la RI (les moteurs de recherche).

automatique des langues, en se plaçant du point de vue de la modélisation des données. Je m'attacherai à montrer, en particulier, que si ces applications ont permis, dans bon nombre de cas, de concevoir des systèmes performants, elles ont aussi souvent conduit à renouveler la compréhension traditionnelle des tâches abordées.

En s'appuyant sur cette typologie, la section 1.3 esquisse une discussion des difficultés auxquelles se confrontent, de manière répétitive, les méthodes statistiques en traitement automatique des langues : la question de la validité des échantillons, de la représentation des données et de l'estimation des modèles. Pour chacune de ces difficultés, j'essaierai de dresser une synthèse des propositions qui ont pu être formulées pour les résoudre et de discuter de leur validité.

Un bilan contrasté de ces travaux conclut ce chapitre (section 1.4) : l'utilisation de méthodes d'apprentissage a indiscutablement contribué au développement d'outils opérationnels de traitement automatique des langues, au détriment parfois d'une réflexion sur les spécificités des données linguistiques et sur les conséquences de ces spécificités pour l'apprentissage de régularités linguistiques[3].

1.1 Introduction

Les domaines du traitement automatique des langues et de l'apprentissage automatique constituent deux branches de l'intelligence artificielle : ils se situent, comme souvent en IA, à la frontière de plusieurs communautés qui expriment des conceptions différentes, parfois antagonistes, de ces champs du savoir. Chacun de ces domaines fait ici l'objet d'une courte présentation, qui servira principalement à poser certains termes et concepts nécessaires aux développements ultérieurs. Au terme de ces présentations, nous discutons des quelques-unes des raisons qui ont conduit ces deux domaines à se rapprocher.

1.1.1 Le traitement automatique des langues

Le domaine du traitement automatique des langues est varié et contrasté : cette présentation se place d'un point de vue suffisamment général pour proposer une vision unifiée des différents types de problème rencontrés dans ce domaine. Dans un premier temps, nous situons rapidement le domaine, avant de présenter quelques applications emblématiques et de caractériser les principales tâches. Une présentation plus détaillée de certaines de ces tâches, en particulier des tâches dites intermédiaires, est fournie en annexe. Des états de l'art récents et complets du traitement automatique des langues se trouvent dans [Dale et al., 2000, Jurafsky and Martin, 2000, Mitkov, 2003] en langue anglaise et dans [Pierrel, 2000] en français.

Quelques lignes de démarcation

Le domaine du traitement automatique des langues s'organise autour des questions de modélisation des actes de production et de perception (reconnaissance, compréhension) des énoncés de langue, en vue de leur accomplissement ou de leur simulation par des machines. Il constitue un point de rencontre entre différents secteurs de la connaissance : linguistique, sciences cognitives, psychologie expérimentale ; il mobilise, adapte et contribue à enrichir, en plus des modèles propres à ces disciplines, des outils empruntés à de multiples domaines : traitement du signal, mathématiques et informatique théorique (logiques, langages formels), intelligence artificielle (représentation des connaissances, apprentissage), etc.

[3] Il est d'ailleurs permis de lire le chapitre 2 de ce manuscrit comme une tentative personnelle pour répondre à cette dernière objection ; alors que les chapitres 3 et 4, plus tournés vers des applications finalisées, s'appuient très directement sur le premier de ces constats.

Informatique linguistique et ingénierie linguistique L'émergence et la constitution progressive de ce champ en tant que domaine de recherche est le résultat historique d'une double évolution [Cori and Léon, 2002]. D'une part, le passage d'une tradition linguistique plus descriptive, visant à enregistrer et à analyser des échantillons de données langagières, à une linguistique plus cognitive, intéressée à comprendre et à modéliser les régularités des langues, et partant, les mécanismes mentaux impliqués dans la production et la perception des énoncés naturels, y compris par le recours à des simulations sur ordinateur. D'autre part, la « révolution cybernétique » vécue par les sociétés industrialisées s'est traduite par une augmentation continue des capacités de stockage et de calcul des ordinateurs et par une numérisation systématique des supports de l'information et de la communication, deux tendances qui se conjuguent pour presser le développement de technologies capables de traiter ces contenus. Cette demande sociale est en réalité première dans l'émergence du traitement automatique des langues comme domaine de recherche, qui se constitue historiquement autour du problème de la traduction automatique. Par ce biais, le traitement automatique des langues rencontre dès ses débuts, les demandes de l'*ingénierie documentaire* et de la gestion électronique des documents.

Par opposition aux travaux plus théoriques, auxquels on associe souvent le terme d'*informatique linguistique* (*Computational Linguistics*), ces travaux plus orientés vers le développement d'applications sont souvent désignés sous le terme d'*ingénierie linguistique* (*Natural Language Engineering*) [Cunningham, 1999, Pierrel, 2000].

Traitement de l'écrit, traitement de l'oral Une seconde ligne de démarcation divise le champ entre l'étude des énoncés oralisés (reconnaissance et synthèse vocale, dialogue oral) et les énoncés écrits. La constitution de deux communautés relativement autonomes partageant un même objet d'étude est le résultat d'une division du travail scientifique fondée sur le type des observations à analyser : du côté du traitement de la parole, les données numériques, correspondant à des échantillons de signal acoustique ; du côté du traitement de l'écrit, les données symboliques, correspondant à des caractères ou des mots imprimés.

Pendant de nombreuses années, chacun de ces domaines s'est développé de manière relativement indépendante, exploitant les outils mathématiques les plus adaptés à chacun de ces types de données : modélisations physiques des organes de phonation et d'audition, méthodes de traitement du signal, modèles statistiques, d'un côté ; modèles traditionnels de l'informatique et de l'intelligence artificielle : grammaires formelles, formalismes de représentation des connaissances, de l'autre.

Les points de rencontre[4] entre ces deux domaines étant devenus plus nombreux au fur et à mesure que les technologies de traitement du signal de parole devenaient plus matures, ces deux communautés ont entamé un rapprochement qui se traduit par une interpénétration croissante des problèmes et des méthodes et par une mise en commun des expériences, au point que certaines présentations générales du traitement automatique des langues ne distinguent plus ces deux domaines (e.g. [Jurafsky and Martin, 2000]).

Applications et ressources

Les tâches du TAL Il est classique de présenter le domaine de l'ingénierie linguistique en l'organisant en grandes tâches, aux entrées/sorties bien identifiées : la traduction automatique, la production de résumés, la génération d'énoncés ou de textes, l'interrogation en langage naturel de bases de données, la synthèse de la parole à partir du texte sont ainsi quelques exemples de ces tâches que l'on appelle *tâches finalisées*. En schématisant grossièrement, ces tâches finales s'organisent en trois types principaux :

– les tâches de production ou d'aide à la production de documents : correction orthographique, génération de texte à partir d'une description formelle, etc. ;

[4] Pratiquement toutes les « grandes » tâches traditionnelles (traduction, résumé, etc.) applicables à des énoncés écrits sont aujourd'hui également envisagées pour des énoncés oraux.

– les tâches liées à la gestion de documents ou de bases documentaires : traduction automatique, résumé, recherche et extraction d'information, etc ;
– les tâches liées à la conception d'interfaces homme-machine : agents dialogant par téléphone, assistants virtuels, etc.

Par opposition, les *tâches intermédiaires* correspondent à des traitements visant à structurer les données dans l'optique de parvenir au degré d'analyse nécessaire à la réalisation d'une tâche finale. L'organisation de ces tâches intermédiaires prend la forme prototypique du *pipe-line*. Ce pipe-line s'organise en grandes « couches » de traitements[5] : segmentation et normalisation, analyse lexicale, syntaxique, sémantique, voire pragmatique de l'énoncé, au cours desquels la représentation (linéaire) initiale de l'énoncé est progressivement enrichie et transformée. D'un flux de caractères, on passe ainsi à un flux de mots, à un arbre d'analyse, à une représentation formelle (formule logique, structure de traits, graphe conceptuel) par enrichissement et structuration progressive de l'entrée, chaque étape étant un prérequis à l'exécution de l'étape de traitement suivante. Dans la suite, j'utiliserai souvent le terme de *tâches d'analyse linguistique* pour faire référence à ces traitements.

Des ressources Il est rapidement apparu que le développement de systèmes complets de traitement des langues ne pouvait s'envisager sans la disponibilité de nombreuses sources de connaissances : règles de segmentation, lexiques, règles morphologiques et syntaxiques, ressources terminologiques et sémantiques. Ces sources de connaissance emmagasinent à la fois une expertise linguistique et une forme de connaissance du monde physique et social.

Les questions liées à la constitution et à la maintenance de ces ressources, aux sources de connaissance autorisées, aux représentations adéquates pour ces connaissances, ont donné lieu à de multiples propositions et débats dans la communauté des informaticiens linguistes. L'unanimité se faisant toutefois pour reconnaître la rareté (et donc la valeur) de ces ressources, au point que cette rareté a longtemps été perçue comme l'obstacle principal à la mise en place d'outils de traitement performants.

Synthèse

Pour la suite de chapitre, retenons principalement que les problèmes du traitement automatique des langues consistent à *enrichir* et à *structurer* des énoncés ou des textes. Cet enrichissement est le plus souvent décomposé en plusieurs étapes, chaque étape donnant lieu au calcul d'une information supplémentaire sur la structure en cours d'élaboration, en s'appuyant sur des sources de connaissance multiples. Chaque étape met donc en jeu une entrée (connue) et une sortie (à construire). Deux caractéristiques des données linguistiques rendent ces calculs difficiles et intéressants :

– les entrées présentent, le plus souvent, une forme d'organisation temporelle, hiérarchique ou relationnelle. Le cas le plus usuel étant celui où elles se présentent sous la forme d'une *séquence de symboles*, l'ordre d'apparition des unités au sein d'un flux étant pertinent (mais pas nécessairement suffisant) pour déterminer leur organisation ;
– les unités qui composent l'entrée sont ambiguës, et ce dans un double sens : d'une part, plusieurs interprétations de l'entrée peuvent être, à un moment donné du traitement, en concurrence : ainsi, dans le cas d'une entrée séquentielle, la segmentation du flux peut n'être qu'imparfaitement connue ; d'autre part, la forme ou la position des unités, prises isolément, ne permettent pas de déduire directement leur rôle dans la structure à construire.

Nous renvoyons, de nouveau, à l'annexe A pour une discussion plus complète des différentes tâches d'analyse linguistique, dans lesquelles s'instancient de multiples manières les deux difficultés évoquées ci-dessus.

[5] Il n'est pas fortuit que cette organisation classique recouvre presque terme pour terme l'organisation modulaire des mécanismes cognitifs proposée par les linguistes, ni qu'elle soit périodiquement remise en cause par les développeurs d'applications finalisées.

1.1.2 L'apprentissage automatique

Le domaine de l'apprentissage automatique s'intéresse aux méthodes inductives permettant d'acquérir des connaissances à partir d'observations d'un phénonème. Cette connaissance peut être exploitée pour des tâches de décision ou de prévision : c'est le cadre de l'apprentissage supervisé ; ou à des fins d'analyse exploratoire ou de structuration d'un ensemble de données : c'est le cadre de l'apprentissage non-supervisé. Nous introduisons rapidement ces deux cadres d'apprentissage en focalisant notre exposé sur les méthodes d'apprentissage statistique utilisées dans un contexte de traitement automatique des langues. Cette brève introduction sera également l'occasion d'introduire quelques unes des notations utilisées dans la suite de ce document. On se reportera par exemple à [Duda et al., 2001, Mitchell, 1997, Cornuéjols and Miclet, 2002] pour des introductions générales aux méthodes de l'apprentissage automatique.

L'apprentissage supervisé

Catégorisation Le cadre général de l'apprentissage supervisé consiste, à partir de l'observation d'un ensemble de données de la forme $\{(\mathbf{x}^{(i)}, y^{(i)}, i = 1 \dots n\}$, à prédire la valeur de y pour de nouvelles valeurs de x. Dans un cadre probabiliste, chaque $\mathbf{x}^{(i)}$ représente une observation d'une variable aléatoire $\mathbf{X}$ multidimensionnelle, prenant typiquement des valeurs dans $\mathbb{R}^m$ (mais ces valeurs peuvent également être discrètes). Il est courant de distinguer deux situations, suivant les valeurs de la variable aléatoire Y : lorsqu'elle prend des valeurs continues, on parle de *régression*, lorsque ses valeurs sont discrètes, de *catégorisation*. C'est uniquement cette seconde situation que nous considérons ici, en nous intéressant particulièrement à la situation où Y est une variable binaire, prenant ses valeurs dans $\mathcal{Y} = \{-1, +1\}$.

Le cadre statistique de la catégorisation supervisée considère le problème de l'apprentissage comme celui de l'induction d'une fonction f de $\mathbb{R}^m$ dans $\{+1, -1\}$. Sous l'hypothèse que les observations sont indépendantes et uniformément distribuées selon une loi de probabilité inconnue P, la meilleure fonction (hypothèse) sera celle qui a une espérance de risque minimum [Vapnik, 1995] :

$$f^* = \operatorname{argmin} E_P[\lambda(f(\mathbf{x}) \neq y)], \tag{1.1}$$

où λ est une fonction de risque (*loss function*), qui évalue le coût induit par une erreur de catégorisation. Le cas le plus simple est celui où $\lambda = \lambda_0$ est une fonction binaire, qui vaut 1 en cas d'erreur et 0 sinon.

La distribution P étant inconnue, ce problème de minimisation n'est pas soluble. À défaut, il est possible de chercher à minimiser le risque associé à une hypothèse f sur les données observées (le *risque empirique*) :

$$f^* = \operatorname{argmin} \frac{1}{n} \sum_{i}^{n} \lambda(f(\mathbf{x}^{(i)}) \neq y^{(i)}), \tag{1.2}$$

Ce cadre général est bien connu et il existe une multiplicité d'outils statistiques pour aborder ce problème : régression logistique, analyse linéaire discriminante, arbres de décisions, k-plus proches voisins, réseaux de neurones, etc. Plus récemment, un effort important a été déployé pour construire et caractériser des stratégies de sélection de l'hypothèse f qui soient explicitement telles que le risque empirique fournisse une approximation raisonnable de l'espérance du risque. Cet effort s'est concrétisé avec le développement de méthodes statistiques puissantes et d'une grande généralité : machines à vecteurs de support[6] (SVMs) [Vapnik, 1995], *boosting* [Freund and Schapire, 1996], modèles de décision probabilistes, etc. Ces modèles se généralisent plus ou moins directement aux situations dans lesquelles Y prend plus de deux valeurs (problèmes multiclasses). Cette question est discutée en particulier dans [Dietterich and Bakiri, 1995, Allwein et al., 2000].

[6] Aussi connus sous le nom de 'Séparateurs à Vaste Marges', qui permet de conserver l'acronyme.

Dans un cadre probabiliste bayésien, le choix de la catégorie d'une nouvelle instance $\mathbf{x}$ est réalisé par la règle du maximum *a posteriori* (MAP).

$$f(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, P(y|\mathbf{x}) \tag{1.3}$$

Cette règle consiste à choisir la classe qui minimise l'espérance du risque λ_0, qui vaut simplement, dans le cas d'une décision binaire, $P(-y|\mathbf{x}) = 1 - P(y|\mathbf{x})$.
Deux approches sont traditionnellement distinguées, selon que le programme de maximisation (1.3) est résolu en modélisant directement $P(Y|\mathbf{X})$ (ces modèles sont dits *discriminants*), ou bien en inversant la dépendance, conduisant à une modélisation de $P(\mathbf{X}, Y)$ (modèles dits *génératifs*, parce qu'ils définissent un modèle de génération des données d'apprentissage).
Les modèles exponentiels sont les modèles probabilistes discriminants les plus utilisés en traitement automatique des langues, exprimant la loi conditionnelle de Y sachant $\mathbf{X}$ sous la forme :

$$P(Y = y|\mathbf{X} = \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\sum_k \theta_k f_k(\mathbf{x}, y)), \tag{1.4}$$

Les θ_k sont les paramètres du modèle (rassemblés dans le vecteur θ) ; chaque f_k est une fonction (le plus souvent une fonction binaire) des observations et des étiquettes de classe et $Z(\mathbf{x})$ opère une normalisation des distributions conditionnelles selon :

$$Z(\mathbf{x}) = \sum_{y \in \mathcal{Y}} \exp(\Sigma_k \theta_k f_k(\mathbf{x}, y))$$

L'estimation des paramètres θ_k est réalisée par exemple par maximisation de la vraisemblance conditionnelle (ou pseudo-vraisemblance) de l'ensemble d'apprentissage.
Les modèles génératifs transforment l'équation (1.3) selon :

$$f(\mathbf{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, P(\mathbf{x}|y)P(y) \tag{1.5}$$

qui fait apparaître deux termes : la vraisemblance de l'observation $\mathbf{x}$ conditionnellement à y et la probabilité *a priori* de y. Dans un cadre supervisé, ces deux distributions sont estimées à partir des observations.

Séquences Un cadre particulier de supervision est celui pour lequel les observations sont des séquences linéairement ordonnées dotées d'une cohérence temporelle : on parle alors de processus. Cette problématique généralise la catégorisation de deux manières différentes.

Une première situation consiste à prédire chaque valeur de Y à partir des observations passées : chaque $\mathbf{x}^{(i)}$ contient les k valeurs précédentes de Y aux « instants » $t-1 \ldots t-k-1$, à partir desquels on cherche à prédire $y^{(i)} = y_t$. Une variante mineure consiste à supposer qu'on observe non pas une mais p séquences distinctes : on considère alors que le le processus de génération est réinitialisé au terme de chaque séquence. Nous désignerons ce cadre par le terme général de *modèles de séquences*.

Une seconde situation correspond au cas où chaque couple d'observations $(x^{(i)}, y^{(i)})$ est lui-même un couple de séquences de même longueur $l^{(i)}$, donnant lieu aux *modèles de transcodage*.

Modèles de séquences Pour ces deux situations, le cadre statistique de référence est celui des chaînes de Markov. Dans la première, le modèle de Markov d'ordre k le plus simple associe à chaque séquence $\mathbf{x}$ de longueur k (il y en a un nombre fini) un état particulier s du modèle. En supposant que la chaîne est initialement dans l'état $s_1 = \mathbf{x}_1 = y_{-k+1} \dots y_0$, la probabilité d'une séquence $y_1 \dots y_l$ est donnée par :

$$P(y_1 \dots y_l) = \prod_{t=1}^{t=l} P(y_t | \mathbf{x_t} = y_{t-m-1} \dots y_{t-1}) \tag{1.6}$$

Chaque séquence est engendrée le long d'un parcours aléatoire dans un graphe d'états, dans lequel chaque état s est uniquement déterminé par les k dernières valeurs de Y. À chaque instant de ce parcours, le nouvel état est choisi selon une distribution de probabilité qui ne dépend que de l'état courant. Dans la mesure où les parcours sont observés, l'estimation de ces modèles peut être réalisée directement par maximisation de la vraisemblance des observations. Des contraintes supplémentaires portant sur la forme des séquences sont introduites en donnant une probabilité nulle à certaines transitions.
Une manière alternative d'envisager la probabilisation d'une séquence est de construire un modèle exponentiel pour l'ensemble de la séquence, selon :

$$P(y) = \frac{1}{Z} \exp(\sum_k \theta_k f_k(y)), \tag{1.7}$$

où chaque f_k est, comme précédemment, une fonction binaire, portant ici sur des propriétés de toute la séquence et où Z opère une normalisation selon (il faut sommer sur l'ensemble $\mathcal{Y}^*$ des séquences de symboles de $\mathcal{Y}$) :

$$Z = \sum_{y \in \mathcal{Y}^*} \exp(\Sigma_k \theta_k f_k(y))$$

Modèles de transcodage Le cadre des modèles de transcodage est un peu différent et présuppose l'introduction d'états cachés, correspondant aux valeurs possibles de Y. Ainsi, les *modèles de Markov cachés* d'ordre 1 décomposent la probabilité d'un couple (x, y) selon[a] :

$$P(\mathbf{x}_1 \dots \mathbf{x}_l, y_1 \dots y_l) = \prod_{t=1}^{t=l} P(y_t | y_{t-1}) P(\mathbf{x}_t | y_t) \tag{1.8}$$

Comme précédemment, chaque couple d'observations est engendré lors du parcours d'un graphe dont chaque état s correspond à une étiquette possible y ; à chaque moment de ce parcours, le modèle émet une valeur pour $\mathbf{X}$ selon $P(\mathbf{X}|s)$ et sélectionne un nouvel état s' (et donc une nouvelle valeur pour Y) selon $P(s'|s)$.
Une introduction détaillée à ces modèles et aux algorithmes permettant d'estimer leurs paramètres se trouve dans [Rabiner and Juang, 1993] (avec un net tropisme pour les applications en traitement de l'oral) et dans [Charniak, 1993] (avec un tropisme pour les applications en traitement de l'écrit).

[a] On suppose que la chaîne est initialisée dans un état initial qui spécifie y_0.

L'apprentissage non-supervisé

Le cadre de l'apprentissage non-supervisé concerne des tâches d'analyse exploratoire des données. Il s'agit, à partir de la seule observation de $\mathbf{x}$, d'extraire, de visualiser et de résumer les corrélations entre des composantes de ce vecteur ; ou bien encore de construire des regroupements homogènes des observations. Ces deux aspects sont rapidement introduits ci-après.

Analyse des corrélations L'analyse des corrélations consiste à rechercher, par des mesures statistiques, des associations positives ou négatives parmi les composantes de $\mathbf{X}$.

Une manière très simple de procéder consiste à mettre en œuvre des mesures statistiques de la corrélation entre les composantes prises deux à deux : dans le cas continu, on utilisera, par exemple, le coefficient de

corrélation empirique, dans le cas discret, la mesure du χ^2 ou encore l'information mutuelle. En notant $\tilde{P}$ les distributions empiriques, ces deux dernières mesures sont définies respectivement par :

$$\chi^2(X_j, X_k) = \sum_{x_j, x_k} \frac{(\tilde{P}(x_j, x_k) - \tilde{P}(x_j)\tilde{P}(x_k))^2}{\tilde{P}(x_j)\tilde{P}(x_k)} \tag{1.9}$$

et par [Cover and Thomas, 1991] :

$$I(X_j, X_k) = \sum_{(x_j, x_k)} \tilde{P}(x_j, x_k) \log(\frac{\tilde{P}(x_j, x_k)}{\tilde{P}(x_j)\tilde{P}(x_k)}) \tag{1.10}$$

Les techniques d'analyse factorielle [Lebart et al., 1979] permettent de rechercher simultanément l'ensemble des corrélations entre les composantes de $\mathbf{X}$. Lorsque les observations $\mathbf{x}$ sont des vecteurs de $\mathbb{R}^m$, *l'analyse en composantes principales* permet de résumer un ensemble de corrélations entre composantes, en construisant le sous-espace d'inertie maximale qui explique au mieux les observations. Cette construction peut être exploitée à des fins diverses : pour réduire la dimension de $\mathbf{X}$ (on ne conserve que les dimensions les plus importantes), ou pour visualiser les données (dans un sous-espace de petite dimension défini par les premières composantes principales). *L'analyse factorielle des correspondances* est une transposition de cette méthode à l'analyse de deux variables discrètes et se généralise à *l'analyse des correspondances multiples* lorsque l'on souhaite étudier simultanément plus de deux variables.

Classification La classification non-supervisée consiste à construire des regroupements[7] homogènes (des classes ou *clusters*) d'objets décrits par un ensemble d'attributs. La constitution de tels agrégats peut être vue, selon les applications visées, comme une fin en soi, ou bien encore comme un moyen d'améliorer des systèmes de catégorisation supervisée en transformant la représentation fournie par $\mathbf{X}$ en une représentation plus simple.

Les méthodes pour opérer de tels regroupements constituent un domaine d'étude riche et foisonnant, qui a donné lieu à de multiples propositions dont le recensement excède de loin le cadre de ce document (voir [Jain et al., 1999] pour un état de l'art récent). Il est néanmoins possible de distinguer trois grandes familles de méthodes :

– les méthodes *vectorielles* se fondent sur une représentation multi-dimensionnelle des objets, représentation dont on peut déduire diverses mesures de similarité. Étant donnée cette similarité, il devient possible de construire des regroupements de manière ascendante, descendante ou encore non-hiérarchique. Ces méthodes conduisent généralement à des affectations booléennes (déterministes) d'un objet à une classe ;
– les méthodes *graphiques*, qui se fondent sur la manipulation d'un graphe pondéré, dans lequel il s'agit, par exemple, de trouver des cliques de taille ou de poids maximum ;
– les méthodes *probabilistes*, qui modélisent les données par un mélange de modèles génératifs. Ces méthodes ont la particularité de permettre que l'assignation d'un objet à une classe soit probabilisée, la probabilité mesurant l'intensité de l'association entre objets et classes. Par ce biais, ces méthodes atteignent un objectif proche de celui visé par les méthodes factorielles, à savoir la construction d'une représentation simplifiée d'un objet sous la forme du vecteur contenant les probabilités d'appartenance aux différentes classes.

[7] Il est également possible de penser que ces algorithmes construisent en fait des sous-populations ou des subdivisions à partir d'une collection hétérogène d'objets.

Classification probabiliste Le modèle probabiliste de référence pour la classification est le modèle de mélange de lois gaussiennes. Dans la variante la plus simple, on fait l'hypothèse du modèle de génération suivant :
– tirage d'une des composantes c du mélange parmi n_C dans une loi discrète $P(C)$.
– tirage d'une valeur pour $\mathbf{X}$ dans une loi gaussienne de paramètres $(\mu_c, \sigma I_m)$.
La probabilité d'une observation $\mathbf{x}$ s'écrit alors :

$$P(\mathbf{x}) = \sum_{c \in \mathcal{C}} P(c) P(\mathbf{x}|c) \tag{1.11}$$

Dans ce modèle, les valeurs de C sont les observations latentes (non-observées). Si la maximisation directe de la vraisemblance des données d'apprentissage n'est pas réalisable, il est en revanche possible de maximiser indirectement la vraisemblance en utilisant l'algorithme EM (voir, par exemple, [Mitchell, 1997]).

Les séquences Le cas où les observations sont séquentielles offre un troisième cadre pour la mise en œuvre de techniques d'apprentissage non-supervisé. Comme pour les tâches supervisées, deux cas sont à distinguer selon que l'on observe une ou plusieurs séquences. Dans le premier cas, l'apprentissage vise à détecter des frontières au sein de la séquence. Dans un cadre probabiliste, ceci est réalisé en construisant un modèle séquentiel des observations (chaque élément de la séquence est prédit par son passé) et en recherchant les points pour lesquels les prédictions du modèle divergent significativement des observations. Dans la suite du document, les modèles réalisant un tel découpage seront désignés sous le nom de *modèles de ruptures*.

Dans le second cas, on suppose que l'on observe un ensemble de séquences, la tâche consistant alors à identifier simultanément (i) une segmentation des séquences en fragments de taille réduite et (ii) une classification des fragments identifiés. Les modèles permettant d'effectuer ce type de traitement seront désignés sous le nom de *modèles de classegment* , néologisme que je propose pour exprimer la double nature de la tâche.

Modèles probabilistes de classegment Lorsque le nombre de classes est connu, ce dernier problème peut également être résolu en utilisant des Modèles de Markov cachés. Pour construire des regroupements en k classes, on supposera de nouveau que chaque séquence est issue d'un parcours dans un graphe à k états (un état par classe), chaque état étant caractérisé par un modèle de génération de $\mathbf{x}$ qui lui est spécifique.

Il est alors possible d'utiliser l'algorithme EM pour déterminer de manière non supervisée une allocation (probabiliste) optimale de chaque élément de la séquence à un état. Cette allocation étant connue, des estimateurs pour les différents paramètres du modèle s'en déduisent sans autre difficulté (voir, de nouveau, [Rabiner and Juang, 1993]).

1.1.3 Une convergence inévitable ?

Il est délicat d'analyser *a posteriori* les raisons qui ont conduit les méthodes isssues de l'apprentissage automatique à prendre la place qu'elles occupent aujourd'hui dans le paysage du traitement automatique des langues. J'ai déjà évoqué plus haut l'influence de domaines voisins, tels que ceux du traitement de la parole et de la recherche d'information, domaines dans lesquels les méthodes statistiques fournissent de nombreux outils. Les méthodes d'apprentissage n'auraient toutefois pu rencontrer un tel succès sans l'existence de dynamiques internes aux domaines du traitement automatique des langues (ou de la linguistique) et de l'apprentissage automatique.

Mentionnons en premier lieu un regain d'intérêt indéniable pour les questions liées à l'apprentissage humain du langage, qui fournit un terrain de rencontre entre ces deux domaines ; mais également un renouveau

des méthodes d'analyse de corpus[8] en linguistique, facilité par la disponibilité croissante d'archives électroniques.

Il est également clair que cette évolution n'aurait pu avoir lieu si, en se confrontant à de nouveaux problèmes et à de nouvelles données, la communauté des apprentistes n'avait su proposer des solutions innovantes, permettant d'améliorer de manière continue les performances des outils disponibles.

Dans cette section, nous revenons rapidement sur ces différents facteurs qui ont, en l'espace d'une grosse décennie, contribué à transformer en profondeur le domaine.

Résurgence de l'empiricisme

Sous l'influence notamment des travaux de Noam Chomsky (qui se diffusent à partir de la fin des années 50), la majorité de la communauté des linguistes s'est pendant un temps désintéressée des méthodes statistiques et de la question de l'apprentissage humain du langage. Ce désintérêt se retrouvant dans une large mesure en informatique linguistique, domaine dans lequel l'essentiel des efforts s'est initialement porté vers l'élaboration de systèmes symboliques à base de règles, s'appuyant sur des systèmes de représentation riches et complexes (structures de traits typées, arbres étiquetés, formules logiques, etc.), et sur la conception d'algorithmes efficaces exploitant ces règles et représentations.

La prise de conscience de l'efficacité des techniques statistiques pour les problèmes de reconnaissance vocale [Jelinek et al., 1975], puis pour des problèmes longtemps considérés comme difficiles, comme la tâche d'étiquetage morpho-syntaxique [Church, 1988, DeRose, 1988] a progressivement conduit à un complet renversement[9] de tendance, au point que les méthodes statistiques occupent aujourd'hui largement le devant de la scène dans les principales conférences internationales du domaine (ACL, COLING).

On peut voir, dans ce retournement de situation, la réhabilitation d'un courant de recherche ancien en linguistique[10] les linguistes (re)découvrant l'intérêt des méthodes statistiques pour étudier et modéliser l'apprentissage humain du langage, la variabilité des usages de la langue, les évolutions linguistiques, ou encore de nombreux phénomènes non-catégoriels, tels que la gradualité des jugements de grammaticalité ou les préférences lexicales. Certains de ces arguments sont développés rapidement ci-après, qui ne doivent toutefois pas occulter le fait que l'utilisation de corpus procède en réalité de positions théoriques et de motivations pratiques très diverses.

En plus des références citées ci-dessous, on pourra également consulter [Church and Mercer, 1993, Abney, 1996, Manning and Schütze, 1999] pour une discussion de cette résurgence de l'empiricisme.

La variabilité Aussi longtemps que l'ambition reste de décrire la connaissance linguistique d'un locuteur adulte particulier, en faisant abstraction des conditions dans lesquels il utilise cette connaissance, il est possible de développer des modèles du langage à base de règles. Lorsque l'on s'intéresse concrètement à des corpus d'énoncés ou de textes, il devient nécessaire de se confronter aux multiples formes de variabilité intra et inter-individuelles, de reconnaître que ce qui est grammaticalement acceptable pour l'un ne l'est pas pour l'autre, que le sens et l'usage des mots varie d'un groupe de locuteurs à un autre, d'un sous-domaine à l'autre, d'une époque à une autre, d'accepter la possibilité que les énoncés soient « bruités », etc. Les besoins

[8] Ici et dans la suite, j'utilise '*corpus*' dans un sens libéral pour désigner l'ensemble des données disponibles pour l'apprentissage. Voir [McEnery and Wilson, 1996, Habert et al., 1997] pour une introduction générale aux linguistiques de corpus.

[9] Ce renversement s'est trouvé grandement facilité par l'existence d'une communauté active, notamment de lexicographes, dont les méthodes ont été réhabilitées, et dont les dictionnaires et corpus ont pu trouver des débouchés nouveaux.

[10] Du point de vue du traitement automatique des langues, ce retournement de situation est également un retour à des sources qui ne se sont jamais taries : historiquement, les travaux en linguistique quantitative et de lexicométrie précèdent (et survivent à) l'arrivée en France des problématiques de traduction automatique, qui déboucheront sur la constitution d'une informatique Linguistique focalisée sur les questions de modélisation logico-algébrique de la syntaxe et de la sémantique [Cori and Léon, 2002].

de traiter des données réelles, pour des applications à large couverture, ont remis ces questions au goût du jour, révélant certaines limites des approches symboliques : leur manque de robustesse, leur difficulté à appréhender les ambiguïtés, les problèmes de portabilité des ressources, etc. Les méthodes statistiques ont l'avantage de fournir un cadre pour aborder ces questions, comme l'illustre la discussion qui suit sur l'ambiguïté.

L'ambiguïté À quelque niveau d'analyse de la langue que l'on se place, les problèmes de désambiguïsation abondent : une même lettre peut se prononcer de plusieurs manières, un même lexème avoir de multiples constructions syntaxiques ou des significations différentes, une même phrase avoir plusieurs analyses, etc Fuchs [1996]. Une observation générale est que cette ambiguïté est grandement sous-estimée par l'humain : la rareté de certaines constructions ou de certains usages suffit à masquer la possibilité d'une interprétation . Face à ces phénomènes, les systèmes symboliques sont placés dans une alternative dont les deux termes sont également mauvais : (i) soit ne pas prendre en compte ces usages rares et minimiser l'ambiguïté réelle des énoncés ; (ii) soit les prendre en compte, conduisant à *exagérer* l'ampleur des ambiguïtés. Par contraste, les modèles probabilistes de la grammaire offrent un cadre pour à la fois prédire l'ambiguïté et pondérer les analyses en compétition, permettant de sélectionner la meilleure des analyses.

Un nouveau terrain d'expérimentation

Il est également indéniable que l'étude de données linguistiques a suscité un fort intérêt[11] et a donné lieu à de nombreuses innovations (voir ci-après, section 1.3) au sein de la communauté de l'apprentissage automatique. Les raisons de cet intérêt sont également multiples, qui tiennent à la fois à la nature particulière des données linguistiques, qui en font un champ d'expérimentation[12] d'une très grande richesse, ainsi, naturellement, qu'à la possibilité de trouver des débouchés pour les techniques d'apprentissage.

Pour ne développer que le premier de ces deux arguments, mentionnons pour commencer que les données linguistiques sont disponibles en très grande quantité. Par rapport aux tâches habituellement considérées dans la communauté de l'apprentissage automatique[13], les problèmes du traitement automatique des langues permettent de manipuler des corpus d'apprentissage contenant des centaines de milliers, voire des millions d'observations. Le déploiement de l'Internet, qui permet d'envisager l'exploitation de sources de données arbitrairement grandes, a accentué cette caractéristique, renouvelant certaines des problématiques du domaine.

Les problèmes de l'analyse linguistique sont ensuite des problèmes difficiles : de nombreuses tâches sont complexes et résistent aux efforts de modélisation. C'est le cas de la traduction automatique, ou encore de la production automatique de résumés, mais même des tâches en apparence plus simples comme la correction orthographique ou l'étiquetage morpho-syntaxique restent imparfaitement résolues et continuent de susciter de nouvelles propositions.

Les données linguistiques sont enfin des données « riches ». En particulier, elles permettent de travailler sur des observations séquentielles exhibant des dépendances (syntaxiques et sémantiques) dont la modélisation est complexe ; sur des problèmes de catégorisation dans un très grand nombre de classes ; et sur des problèmes en très grande dimension, utilisant des descriptions à plusieurs niveaux des énoncés ou des textes. Cette richesse du domaine se mesure également à la quantité des bases de connaissance (lexiques, corpus

[11] Cet intérêt est attesté par les nombreux travaux consacrés à des expérimentations conduites sur des données textuelles dans les principales conférences du domaine (ICML, ECML, NIPS, etc.).

[12] Les mêmes raisons valent sans doute pour la diffusion des méthodes statistiques en traitement des données biologiques et bio-médicales.

[13] Voir, par exemple, la base de données de l'UCI, qui regroupe un ensemble de jeu de données de référence pour évaluer de nouveaux apprentis [Hettich et al., 1998].

annotés, ressources terminologiques et sémantiques, etc.) disponibles sur les problèmes de l'analyse linguistique. Ce domaine ouvre donc un large champ d'expérimentation aux méthodes d'apprentissage cherchant à intégrer des connaissances *a priori* du domaine.

Dernier facteur facilitant, probablement pas le moindre : le traitement automatique des langues est un domaine relativement ouvert, au sens où la compréhension des tâches à réaliser ne nécessite que rarement une connaissance fine des théories linguistiques sous-jacentes et où tout locuteur (apprentiste ou non) peut relativement facilement comprendre et s'approprier les problèmes posés.

1.1.4 Conclusion

Dans cette section, nous avons rapidement introduit les domaines du traitement automatique des langues et de l'apprentissage automatique, en présentant successivement les principaux traitements impliqués dans les tâches de traitement automatique des langues ; les principaux outils (en particulier statistiques) de l'apprentissage automatique ; puis en discutant quelques-unes des raisons théoriques et historiques qui ont progressivement conduit à la prédominance actuelle des méthodes statistiques pour réaliser de nombreuses tâches d'analyse linguistique.

1.2 Une typologie des applications de l'apprentissage en traitement automatique des langues

La rencontre des deux domaines de recherche présentés dans la section précédente a donné lieu à un foisonnement de travaux, dont il paraît délicat de dégager une vue d'ensemble. C'est pourtant l'objectif de cette section, qui vise à identifier rétrospectivement les principales applications de méthodes d'apprentissage automatique en traitement automatique des langues et à en présenter une vision cohérente.

Le point de vue que je défends ici est que les succès des méthodes utilisant des outils d'apprentissage reposent plus sur des choix appropriés de représentation et de reformulation des problèmes d'analyse linguistiques qui les rendent susceptibles de donner lieu à la mise en œuvre de méthodes connues et bien maîtrisées que sur une adaptation raisonnée des outils de l'apprentissage automatique aux spécificités des données linguistiques.

Comme j'essaierai de le montrer, brosser ce panorama est donc plus aisé qu'il ne semble, dans la mesure où, si les techniques d'apprentissage automatique ont été appliquées pour aborder un grand nombre de tâches et de problèmes, les *reformulations* de ces problèmes se ramènent finalement aux quelques grands types introduits à la section 1.1.2 : identification non-supervisée d'associations et de regroupements, catégorisation supervisée et modélisation des séquences.

Cette section débute donc par une présentation de diverses applications de méthodes d'apprentissage non-supervisé dans un cadre de traitement automatique des langues, qui visent toutes à répondre à la problématique de l'*acquisition automatique de ressources*.

Les trois sections suivantes présentent des contributions utilisant les techniques d'apprentissage supervisé, que cette utilisation soit une mise en œuvre directe d'outils de catégorisation (section 1.2.2) ; qu'elle exploite des modèles de séquence (section 1.2.3) ; ou bien qu'elle repose sur des reformulations idoines de la tâche (section 1.2.4). Que le lecteur ne soit pas surpris, ni trompé, par le nombre délibérément réduit de références à la littérature : chaque sous-problème a fait l'objet d'une multitude d'expérimentations et de comparaisons empiriques, donnant lieu à une littérature abondante (voir, par exemple, la bibliographie de [Lluís Màrquez, 2000]). En ce qui concerne les tâches que j'ai étudiées, on trouvera des références plus complètes dans les trois chapitres qui suivent. Pour les autres tâches, nous renvoyons à l'annexe A, qui décrit plus précisément

les modules des systèmes de traitement automatique et inclut des références complémentaires aux travaux utilisant des méthodes d'apprentissage.

Avant de débuter ce tour d'horizon, il est juste de signaler également certaines omissions, dont la présentation détaillée aurait considérablement allongé ce texte : en particulier les multiples modèles proposés pour probabiliser des langages hors-contexte ou faiblement sensibles au contexte (voir, par exemple, [Schabes, 1992, Charniak, 1993, Bod, 1995, Collins, 1999]) dans le cadre d'applications à l'analyse syntaxique ; ou encore ceux qui furent élaborés dans le cadre de la traduction automatique (en particulier [Brown et al., 1990, Och and Ney, 2002]). Faute d'avoir directement travaillé sur ces questions, nous ne rendrons pas compte ici de ces propositions, qui figurent pourtant parmi les plus innovantes des contributions récentes au traitement statistique des langues.

1.2.1 Les problèmes d'apprentissage non-supervisé

L'apprentissage non-supervisé recouvre deux grands types de techniques : la construction de classes et la détection de corrélations. Diverses applications de ces techniques à la problématique de l'acquisition de ressources sont présentées dans ce paragraphe, ces ressources pouvant être destinées à une validation humaine ou qu'elles soient requises pour mettre en œuvre d'autres outils de traitement.

Classification

Un premier ensemble de problèmes ressortit directement au cadre de la classification non-supervisée :
- constitution de classes de phonèmes dans un flux phonétique (version simple d'acquisition de la phonologie) par analyse de leur contexte d'occurrence ;
- identification de classes lexicales sur la base de leurs contextes syntaxiques (e.g. [Finch and Chater, 1992, Brown et al., 1992]) ;
- identification de relations sémantiques entre formes ou de classes de formes sémantiquement apparentées [Hindle, 1990, Pereira et al., 1993, Rooth, 1995] ;
- identification de groupes de documents thématiquement homogènes dans une collection de documents (voir [Hofmann, 2001, Nigam et al., 2000, Blei et al., 2002] pour des travaux récents, mais cette tradition est bien plus ancienne, voir par exemple [Benzécri and ali., 1981] ainsi que de nombreux travaux en recherche d'information) ;

Pour chacune de ces tâches, la construction de regroupements peut être réalisée en utilisant les outils existants, pour autant qu'on sache :
- définir une représentation multi-dimensionnelle pour les unités qui font l'objet des regroupements : pour les classes syntaxiques cette description intègre, par exemple, les fréquences d'occurrence des formes voisines ; pour les classes sémantiques, le voisinage est étendu à des fenêtres plus larges et n'intègre que des mots sémantiquement pleins ; pour la classification de corpus, elle se fonde sur les fréquences des formes dans le document.
- spécifier une mesure de similarité ou de distance dans l'espace de représentation si l'on choisit un modèle de clustering vectoriel ou spécifier un modèle de génération pour les données, si l'on a recours à un modèle probabiliste.

La classification obtenue fournit des ressources (classes phonétiques, syntaxiques, sémantiques, groupes de documents) directement exploitables en tant que telles ou bien encore disponibles pour d'autres types de traitements. En particulier, la construction de classes de mots a été utilisée de manière répétitive comme une réponse possible au problème de la rareté des occurrences de formes et aux problèmes d'estimation induits par cette rareté (voir, sur ce dernier point, la section 1.3.1). De même, la construction de classes de documents peut servir de préalable à la constitution de modèles statistiques de langage adaptés thématiquement [Seymore and Rosenfeld, 1997].

Détection de ruptures et classegment

Le traitement d'événements séquentiellement ordonnés est au centre de nombreux problèmes du traitement automatique des langues. Les modèles de rupture (cf. la section 1.1.2) ont donc un champ d'application direct et ont été notamment étudiés dans le contexte de l'identification de frontières thématiques dans un flux documentaire [Hearst, 1997, Reynar, 1998, Choi, 2000].

Une situation plus commune en traitement automatique des langues implique toutefois la résolution simultanée des problèmes de segmentation et de classification d'un ensemble de séquences, tâches que nous avons désignées sous le nom de classegment. Ainsi, par exemple :
– apprentissage non-supervisé de la morphologie, vu comme l'inférence simultanée de classes de morphèmes et d'une segmentation en morphèmes d'un ensemble de formes [de Marcken, 1996, Gaussier, 1999, Goldsmith, 2001] ;
– segmentation d'un document en fragments thématiquement homogènes, demandant d'identifier à la fois un ensemble de thèmes et une segmentation thématique du document [Hearst, 1997].

Analyse des co-occurrences

L'analyse de co-occurrences fournit un troisième cadre pour aborder de nombreuses questions portant sur de l'acquisition automatique de ressources. Ces techniques trouvent en effet un champ d'application pour l'acquisition non-supervisée de collocations [Church and Hanks, 1991, Smadja, 1993] ou de listes de termes potentiels (voir, par exemple, [Daille, 1994], qui compare diverses mesures d'association).

Dans ce contexte, la question posée revient essentiellement à mesurer statistiquement l'association de deux formes en examinant leurs occurrences dans un corpus. En associant une variable aléatoire booléenne à chaque forme correspondant à l'occurrence d'une forme dans une phrase (ou dans un syntagme, ou dans une fenêtre glissante, etc.), l'association statistique entre deux formes se mesure directement par la corrélation entre ces deux variables.

1.2.2 Les problèmes de catégorisation

Si les problèmes de l'acquisition automatique de ressources sont plutôt modélisés par les outils de l'apprentissage non-supervisé, le cadre de la catégorisation fournit un second cadre permettant de modéliser un grand nombre de problèmes, en particulier des problèmes de désambiguïsation.

Les différentes tâches de catégorisation de documents (identification de langue, routage, filtrage) [Yang, 1997, Sebastiani, 2002] fournissent autant d'exemples d'applications finales utilisant directement ces techniques. La tâche de désambiguïsation sémantique est un exemple de tâche intermédiaire qui se modélise également sous cette forme [Gale et al., 1992]. Si les tâches qui sont « purement » des problèmes de classification sont assez rares, le champ d'application de ces outils excède en fait largement les deux exemples précédents. Il intègre, en effet, toutes les tâches qui consistent soit à segmenter une séquence, soit, plus généralement, à affecter une étiquette parmi un ensemble fini à chaque position d'une séquence.

Segmenter Les problèmes de segmentation supervisée correspondent à l'identification automatique de frontières dans une séquence. Cette tâche peut être envisagée comme l'assignation d'une variable booléenne aux éléments de la séquence, 1 marquant conventionnellement le début d'un segment. La segmentation peut être syntaxique : il s'agit de distinguer des segments jouissant d'une certaine forme d'autonomie vis-à-vis des segments adjacents ; ou bien sémantique : il s'agit alors d'identifier des segments qui jouent un rôle distingué au sein de la séquence.

Cette tâche s'instancie de multiples manières, comme le montrent les quelques exemples suivants :
– segmentation d'un flux de symboles orthographiques[14] ou de symboles phonétiques en syllabes ou en mots ;
– segmentation d'un flux de caractères en graphèmes ou en « morphèmes » [van den Bosch et al., 1996] ;
– segmentation d'un flux de caractères en phrases [Reynar and Ratnaparkhi, 1997] ;
– segmentations d'une séquence de mots ou d'étiquettes morpho-syntaxiques en syntagmes non-récursifs (*chunks*) ou encore en mots prosodiques ou phonologiques [Ramshaw and Marcus, 1995, Zhang et al., 2002] ;
– segmentation *supervisée* d'un flux documentaire en blocs thématiquement homogènes [Beeferman et al., 1999].
Suivant les tâches, le degré d'ambiguïté des éléments de la séquence peut être inégalement distribué : ainsi dans la modélisation classique de la tâche de segmentation en phrases, pour laquelle seuls les symboles de ponctuation sont ambigus[15] ; ainsi encore la détection des mots accentués dans les systèmes de synthèse vocale, qui ne s'intéresse qu'aux mots sémantiquement pleins.

Étiqueter Les problèmes d'étiquetage correspondent au cas où le nombre de catégories à identifier est supérieur à deux ; toutes les catégories n'étant pas nécessairement possibles pour tous les symboles du flux. On retrouve, dans cette catégorie, une série de tâches classiques :
– la conversion graphème-phonème : les objets catégorisés sont des lettres, auxquels on assigne des catégories phonémiques [Sejnowski and Rosenberg, 1987, Daelemans et al., 1993, Yvon, 1996] ;
– l'étiquetage morpho-syntaxique : les catégories assignées aux formes sont des étiquettes morpho-syntaxiques [Brill, 1993, Daelemans et al., 1996] ;
– la repérage des entités nommées : des catégories différentes sont attachées aux formes qui débutent, continuent, ou sont à l'extérieur pour un type d'entité (on se fera une idée du spectre des méthodes proposées pour cette tâche en consultant par exemple [Tjong Kim Sang, 2002]) ;
Pour des tâches de segmentation et d'étiquetage, entraîner un système de catégorisation est donc une option possible, dont la mise en application ne demande que (i) la disponibilité de données d'entraînement dans lesquelles les flux d'entrée et de sortie sont alignés et (ii) la conception d'un ensemble de descripteurs adéquats. Les descripteurs peuvent porter sur les propriétés intrinsèques de l'objet à étiqueter ainsi que sur son contexte d'occurrence, modélisé par une fenêtre de taille fixe autour de l'unité courante.

Ce cadre permet également, dans une certaine mesure, de prendre en compte des dépendances entre éléments de la séquence de sortie. [Daelemans and van den Bosch, 2005, chap. 7] discute longuement ces »astuces« dans le cadre général des apprentis de type k-plus-proches voisins :
– une première manière de procéder consiste à intégrer l'étiquette prédite à »l'instant« t dans l'ensemble de descripteurs disponibles à l'étape $t-1$;
– une alternative, baptisée *stacking*, consiste à empile des classifieurs à la façon d'un pipe-line : le classifieur de niveau k utilisera comme descripteurs des fenêtres de contexte *intégrant toutes les sorties de ses* $k-1$ *précédecesseurs* ;
– une troisième manière de faire consiste à prédire non pas des symboles isolés, mais des blocs de symboles se recouvrant (ainsi, pour la tâche de étiquetage morpho-syntaxique, chaque classe prédite correspondra à un trigramme d'étiquettes).

1.2.3 Les tâches « séquentielles »

Pour les tâches d'étiquetage, l'alternative consiste à mettre en œuvre des modèles qui sont explicitement conçus pour des dépendances entre unités adjacentes dans un flux, modèles que nous avons introduits à

[14] Pour les langues qui ne marquent pas explicitement cette séparation, comme le chinois ou le thaï.
[15] Au sens où ce sont les seuls qui sont, suivant les contexte, en fin de phrase (étiquetés 1) ou en position interne (étiquetés 0).

la section 1.1.2 sous le nom de modèles de transcodage. De multiples expérimentations ont été également conduites à l'aide de ces modèles, en particulier en utilisant des modèles de Markov cachés ou l'un de leurs nombreux avatars (modèles de Markov du maximum d'entropie [McCallum et al., 2000, Punyakanok and Roth, 2001], champs conditionnels aléatoires [Lafferty et al., 2001], réseaux bayésiens dynamiques [Murphy, 2002], etc).

Les modèles de séquences markoviens ont rencontré une popularité équivalente, sinon supérieure, en particulier dans le contexte de la modélisation linguistique en reconnaissance vocale (voir par exemple [Jelinek, 1997]). Les *modèles de langage* ressortissent en effet presque directement au cadre défini à la section 1.1.2, la seule différence étant que la variable Y prend ses valeurs non pas dans $\{+1, -1\}$, mais dans un ensemble de formes orthographiques potentiellement grand (comprenant typiquement plusieurs dizaines de milliers d'éléments). Ce changement d'échelle pose, en soi, des problèmes difficiles et son étude a donné lieu à une vaste littérature (par exemple [Goodman, 2001, Rosenfeld, 2000]).

1.2.4 Problèmes dérivés

Les tâches du traitement automatique des langues ne sont pas toutes susceptibles de se prêter aussi directement à la mise en œuvre de systèmes d'apprentissage automatique. Par exemple, pour envisager l'utilisation de modèles de transcodage, il faudra que la tâche soit telle que : (i) l'opération d'analyse consiste en l'assignation d'une étiquette catégorielle prenant des valeurs dans un ensemble fini ; (ii) chaque symbole du flux reçoit une (ou plusieurs) étiquette(s) dans un ensemble fini ; (iii) l'organisation linéaire de la sortie reflète celle de l'entrée. Pour prendre deux exemples importants, les tâches d'analyse syntaxique et de traduction automatique ne rentrent pas dans ce cadre. Pour la première, il s'agit de construire un appariement entre un flux d'entrée et une structure récursive (arbre de dérivation ou structure de traits), dont la complexité dépasse les structures linéaires considérées jusqu'à présent. Pour la seconde, segmentation et classification doivent avoir lieu simultanément ; l'ordre et le nombre des « étiquettes » de sortie ne respectent pas nécessairement ceux des étiquettes d'entrée[16], nécessitant de conjoindre au système d'étiquetage des modèles d'alignement plus riches.

Pour étendre le champ d'application des techniques statistiques à d'autres formes de problèmes, deux réponses principales ont été proposées. La première consiste à *transformer* les tâches, soit en les simplifiant, soit en les découpant en une série de sous-problèmes traitables dans un cadre de catégorisation, soit en les transformant en un problème de réordonnancement. Les paragraphes suivants détaillent ces trois approches.

Simplification Une première simplification rendue quasi-nécessaire par l'utilisation d'outils statistiques (y compris ceux non évoqués ci-dessus), consiste à minimiser l'ambiguïté. Alors qu'il existe des variantes dans la prononciation d'une phrase, qu'il existe des phrases syntaxiquement ambiguës, ou sémantiquement polysémiques, et des documents multi-catégoriels, les modèles considèrent, dans leur immense majorité, chaque tâche comme possédant une sortie déterministe. Si, le plus souvent, cette approximation est bénigne et se corrige simplement (par production de sorties ambiguës), notons qu'elle ne l'est pas toujours (cf. la conversion graphème-phonème, pour ne pas parler de la traduction automatique).

Autre exemple d'une telle simplification : [Turney, 2006] envisage la qualification de la relation sémantique entre un nom et son modifieur comme un problème de catégorisation dans 30 classes[17], chaque classe correspondant à une relation sémantique.

Une simplification beaucoup plus drastique est opérée en tâche de résumé automatique : une conception répandue (le *résumé par extraction*) l'envisage comme la sélection d'un sous-ensemble de phrases pertinentes

[16] C'est en fait également le cas en conversion graphème-phonème ; toutefois, les cas dans lesquels un réordonnancement est nécessaire sont suffisament rares pour qu'on puisse les ignorer et se ramener à une tâche d'étiquetage.

[17] Le problème s'avérant encore trop difficile, il considère ensuite, avec de meilleurs résultats, la catégorisation dans 5 classes.

d'un document. La tâche devient alors de discriminer les phrases pertinentes, sur la base d'un ensemble d'attributs contenant, par exemple, leur position dans le texte, dans le paragraphe, le nombre de mots nouveaux qu'elles apportent, etc. [Kupiec et al., 1995].

[Lewis, 1991] utilise une démarche similaire dans un cadre d'extraction d'information : repérant dans le formulaire à remplir un ensemble de champs pour lesquels les valeurs possibles sont en nombre fini, il en déduit un ensemble de catégories, chaque catégorie correspondant à une combinaison particulière de ces valeurs. Comme le montre Lewis, le problème d'extraction peut ainsi être reformulé comme un simple problème de catégorisation supervisée.

Sous-modularisation La tâche d'analyse syntaxique est un exemple de problème, qui, faute de pouvoir directement être abordé comme un problème de catégorisation, a fait l'objet d'une *sous-modularisation*. Cette démarche consiste à envisager un processus d'analyse complexe comme un processus incrémental, au cours duquel la structure se construit par enrichissement progressif de l'entrée. Le point crucial est que chaque étape de ce processus peut être modélisée sous la forme d'un problème de catégorisation ou de balisage : étiquetage morpho-syntaxique, segmentation en syntagmes non-récursifs (*chunks*), désambiguïsation des rattachements et des relations de dépendance, s'enchaînent ainsi pour construire des analyses aussi riches que celles que fournirait une analyse syntaxique classique.

Autre exemple d'une telle sous-modularisation, celui du calcul de la structure prosodique d'un énoncé, qui implique successivement de segmenter en groupes prosodiques (localisation des divers types de pauses), d'identifier les mots focalisés, de segmenter le flux de phonèmes en syllabes, d'identifier les syllabes accentuées, toutes étapes qui, nous l'avons noté, sont réalisables par des méthodes standard de catégorisation.

La constitution de la tâche de repérage des entités nommées comme une sous-tâche particulière du traitement automatique des langues, faisant suite aux premières expérimentations en extraction d'information, fournit un troisième exemple de sous-problème qui a progressivement émergé dans le cadre du découpage d'une tâche finale plus complexe.

Réordonnancement Le réordonnancement (*reranking*) utilise une stratégie différente, particulièrement bien adaptée aux problèmes de traitement de séquences. L'idée générale consiste à coupler un *générateur*, qui propose un ensemble d'analyses possibles pour un flux d'entrée, avec un *évaluateur*, qui ordonne les structures proposées. Le couplage du générateur et de l'évaluateur fournit un système associant à chaque entrée une sortie optimale (la mieux classée par l'évaluateur). Cette stratégie a l'avantage de s'appliquer à des entrées et des sorties très variées. Illustrons sa mise en œuvre sur une tâche de réaccentuation ou plus généralement de correction orthographique[18] qui impliquerait les deux étapes suivantes :

– **Génération :** construction, par accès lexical, de l'ensemble de toutes[19] les versions accentuées d'une phrase non-accentuée.
– **Évaluation :** ordonnancement des phrases en fonction de la probabilité que leur donne un modèle probabiliste de séquences. La correction proposée est celle qui est la mieux « notée » par l'évaluateur.

Cette approche sera naturellement d'autant plus efficace que l'évaluation privilégie les « bons » candidats. Une manière de s'en assurer consiste à *optimiser*, par apprentissage, les scores de l'évaluateur. À partir de l'observation de couples entrée/sortie, l'évaluateur va progressivement améliorer sa manière de réordonner les candidats de façon à faire en sorte que la sortie correcte apparaisse le plus souvent avec un bon rang de classement.

Cette idée a été utilisée en reconnaissance vocale, pour reclasser les transcriptions issues d'une première passe de décodage acoustique [Roark et al., 2004]. Elle a également été utilisée pour des tâches d'analyse

[18] La reconnaissance vocale, lorsqu'elle utilise plusieurs passes de traitements, repose sur un principe similaire.

[19] Il n'est pas toujours nécessaire de considérer toutes les sorties possibles, un premier filtrage pouvant être opéré lors de l'étape de génération.

syntaxique : le modèle de génération est une grammaire hors-contexte probabiliste, qui propose un ensemble d'arbres d'analyse ; le modèle utilisé pour l'évaluation est un modèle de séquences dont les paramètres sont optimisés durant l'apprentissage pour minimiser les erreurs de réordonnancement [Collins and Duffy, 2002] ; ou encore pour des tâches de question-réponse : le module de génération est un moteur de recherche qui propose des réponses possibles ; le module d'évaluation est un modèle probabiliste exponentiel entraîné pour bien classer les réponses correctes [Ravichandran et al., 2003].

1.2.5 Conclusion

Dans cette section, nous avons tenté de donner une vue unifiée des approches à base d'apprentissage pour des problèmes de traitement automatique des langues, en considérant successivement les approches, principalement non-supervisées, qui portent sur la constitution de ressources ; puis celles qui s'attaquent à des problèmes de décision et de résolution d'ambiguïté, le plus souvent formalisés par des méthodes d'apprentissage supervisé.

Au delà de la diversité des tâches étudiées, ces méthodes partagent de nombreux fondements communs et s'appuient le plus souvent sur des descriptions de surface des entités linguistiques, à partir desquelles des décomptes d'occurrences d'événements discrets sont opérés. Il devient dès lors possible d'identifier un certain nombre de problèmes généraux qui se posent à toutes ces approches. Ces problèmes sont discutés dans la section suivante.

1.3 Des questions difficiles

Le rapide survol du domaine effectué à la section précédente pourrait laisser l'impression trompeuse qu'il aura suffi aux concepteurs de ressources linguistiques ou de systèmes de traitement automatique des langues d'ouvrir la boîte à outils statistiques pour y trouver, prêtes à l'emploi, des méthodes bien adaptées à leurs problèmes.

Cette section vise à corriger cette impression, en montrant qu'un certain nombre de spécificités des données et des problèmes d'analyse linguistiques font qu'ils résistent à la modélisation statistique traditionnelle. En réponse à ces résistances, il aura fallu proposer des solutions originales pour atteindre des niveaux de performance satisfaisants et s'attaquer à des problèmes difficiles, dont beaucoup font encore l'objet d'une recherche active.

Dans cette section, nous discutons de ces points d'achoppement, en considérant successivement la question des échantillons de données utilisés pour l'apprentissage, les questions liées à la représentation de ces données, enfin les problèmes de construction et d'estimation de modèles pour ces représentations.

Au terme de cette discusssion, je conclus cette section par l'identification de quelques thèmes de recherche qui me semblent s'inscrire dans le prolongement naturel des travaux actuels en apprentissage statistique.

1.3.1 Des données

Une première problématique commune à tous les modèles à base de données concerne la question des corpus d'apprentissage et de leur représentativité. Dans le cadre des modèles probabilistes, l'hypothèse traditionnelle est que les observations constituant le corpus sont *indépendantes* et tirées suivant une même loi (inconnue). Les modèles sont estimés dans le but de généraliser correctement à de nouvelles observations tirées selon *la même loi*. Dans le cas de données linguistiques, chacune de ces deux hypothèses pose problème et doit être évaluée avec soin.

Concernant la première de ces deux questions, notons simplement pour l'instant que la prise en compte de dépendances entre des observations au sein d'une même séquence, voire au sein d'un même document, a conduit à l'adoption de modèles spécifiques, qui ont progressivement supplanté les modèles de catégorisation pour les tâches d'étiquetage (c.f. la section 1.2.3). Nous reviendrons sur cette question, sur laquelle il reste beaucoup de travail à accomplir, dans la section 1.3.4.

La seconde question renvoie à un problème difficile, qui reste relativement peu abordé (excepté, sans doute, dans la communauté des lexicographes), celle de l'échantillonnage et de la *représentativité* des données linguistiques disponibles lors de l'apprentissage. La réalisation d'une grande diversité des usages, des genres et des registres d'utilisation de la langue n'est pas nouvelle. C'est, de longue date, un des arguments qui fonde les approches empiriques en linguistique. Les travaux en linguistique de corpus, notamment ceux de Biber (par exemple [Biber, 1993]), ont ainsi clairement mis en évidence les fortes variations dans les distributions d'événements linguistiques à travers les genres et les registres textuels. À genre et registre donnés, la variabilité inter-auteurs est même encore telle qu'elle fonde la possibilité de mettre en œuvre de techniques statistiques de stylométrie.

Si les modèles statistiques fournissent un cadre dans lequel des solutions opérationnelles à ces questions ont pu être proposées, en particulier autour de la problématique de l'adaptation, il n'est toutefois pas certain que toutes les conséquences en aient été tirées. Il semble ainsi admis, dans la communauté, que la piste la plus sûre pour améliorer les performances des modèles statistiques consiste à d'augmenter la taille des corpus utilisés : disposer de plus de données permettra d'obtenir des estimations plus fiables (moins variables) des paramètres. Cela est indiscutable, du moins tant que la source de production des données ne change pas et que les données sont homogènes. Cette dernière hypothèse est souvent prise en défaut : l'analyse de plus gros corpus fait toujours émerger son lot de mots, de sens et de constructions syntaxiques « inconnus », qu'il faudra intégrer dans le modèle, en augmentant le nombre de paramètres. Dans ce contexte, il est même envisageable que l'augmentation des volumes de données traités conduise paradoxalement à une dégradation de la qualité des estimateurs[20].

Adaptation

La question de l'adaptation des modèles statistiques à des données nouvelles, différentes des données d'apprentissage, a ainsi fait l'objet de nombreux travaux dans le contexte de la reconnaissance vocale (voir, par exemple, [Lau et al., 1993, Rosenfeld, 1994, DeMori and Federico, 1998, Clarkson, 1999], pour ne parler que de l'adaptation des modèles linguistiques). Pour cette application, il est en effet courant de devoir construire des modèles de langage pour l'oral à partir de données écrites ; ou encore d'estimer des modèles à partir d'un petit ensemble de transcriptions. La question se pose même lorsque l'on considère un registre linguistique bien défini, comme « style » journalistique, ne serait-ce que parce que les thèmes de l'actualité de demain ne sont pas (en même proportion) les mêmes que ceux de l'actualité d'hier, que de nouveaux descripteurs (typiquement de nouvelles formes) apparaissent quotidiennement, quand d'autres voient leur probabilité d'occurrence décroître. Pour l'essentiel, les techniques d'adaptation se répartissent en deux familles :

– l'adaptation des paramètres, qui consiste à déformer, à partir d'un petit ensemble d'apprentissage (dit corpus d'adaptation), des distributions de probabilités apprises sur des corpus généraux ;
– l'adaptation des données, qui consiste à sélectionner dans un gros corpus d'apprentissage les échantillons les plus semblables à ceux disponibles dans le corpus d'adaptation.

[20] Ce paradoxe n'est pas sans rappeler celui que l'on observe en modélisation syntaxique : réduire l'ambiguïté impose de manipuler des descriptions plus précises ; mais ces descriptions plus précises ont également pour effet d'augmenter le nombre d'ambiguïté qu'il faudra résoudre.

Adaptation de modèles probabilistes Du point de vue technique, une solution proposée dans le cadre des modèles de langage exponentiels consiste à utiliser des estimations « robustes », obtenues sur un gros corpus, comme des *a priori* sur les paramètres à adapter. Pratiquement, cett démarche s'implémente simplement lorsque l'on considère des distributions de probabilité ayant la forme :

$$P(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} P_0(y|\mathbf{x}) \exp(\sum_k \theta_k f_k(\mathbf{x}, y)), \tag{1.12}$$

(à comparer avec l'équation (1.4)), où :
– P_0 représente une distribution générale, dont l'estimation robuste est effectuée sur de très grosses collections de données, en utilisant des contraintes aussi complexes que possible ;
– les paramètres du modèle exponentiel sont ceux pour lesquels les estimations sur le corpus d'adaptation sont fiables : on utilisera seulement, par exemple, des fonctions f_k qui repèrent des occurrences d'unigrammes.

Ces modèles sont connus sous le nom de MDI (*Minimum Discrimination Information*) car leur solution est la distribution la plus proche de P_0 (au sens de la divergence de Kullback-Liebler) respectant les contraintes du corpus d'adaptation.

Hétérogénéité et homogénéité de la langue

Pour les modèles statistiques, cette hétérogénéité intrinsèque des données linguistiques implique une dégradation des performances lorsque la nature des données change [Sekine, 1997].

Le cadre méthodologique de référence consiste toutefois à partir d'un corpus homogène, aléatoirement partitionné en ensemble d'apprentissage et ensemble de test, sur lequel les performances en généralisation sont mesurées. Cette pratique présente deux inconvénients majeurs :

(i) elle conduit à des estimations biaisées des paramètres du modèle et à une surévaluation de leurs performances.
Ainsi, en catégorisation de documents, la contribution des termes des corpus d'entraînement sera fortement surévaluée, alors que les termes absents de ce corpus ne seront pas modélisés et devront être ignorés lors de l'exploitation du modèle.

(ii) elle traite chaque tâche comme radicalement différente de la précédente, l'apprentissage étant repris à zéro pour chaque nouveau corpus. Cette pratique conduit donc à laisser de côté des sources d'information potentiellement utiles, dans la mesure où l'homogénéité d'une langue se traduit aussi par des régularités (en particulier syntaxiques) qui transcendent genre et registre.

Les tentatives pour remédier à (ii) restent rares[21], hormis en modélisation du langage : une heuristique communément utilisée consiste ainsi à combiner linéairement des modèles appris sur des ressources hétérogènes, puis à optimiser les poids de la combinaison sur un corpus de développement.

Quant à (i), la réponse majoritaire semble consister à augmenter indéfiniment la taille des corpus : comme l'écrivent Church et Mercer [1993] (repris dans [Church, 2003])

> Thus, it would appear that "more data are better data", (...)

Cette stratégie semble aujourd'hui d'autant plus « payante » que (i) la taille des données disponibles continue d'augmenter et que (ii) compte-tenu de leur taille, les modèles statistiques utilisés en traitement automatique des langues sont souvent éloignés de leurs performances asymptotiques [Banko and Brill, 2001] (voir également ci-après).

[21] Le récent travail de [Daume III and Marcu, 2006], qui utilise un modèle de mélange pour rendre compte de l'hétérogénéité des données disponibles, est un bon exemple d'une démarche probabiliste pour aborder ces questions.

Données non-étiquetées

S'il est souhaitable de chercher à disposer de plus de données, l'annotation nécessaire à l'emploi de méthodes d'apprentissage a un coût non-négligeable. Cette constatation a suscité de nombreux travaux autour de l'exploitation de de données non-étiquetées, qui, avec notamment l'explosion de l'Internet, sont disponibles dans des quantités croissantes.

L'outil statistique de référence pour mettre en œuvre cette stratégie est l'algorithme EM. Dans le cadre du traitement automatique des langues, la grande dimensionalité de l'espace des paramètres multiplie les optima locaux, qui sont autant de points vers lequel l'algorithme peut converger. Son utilisation en mode entièrement non-supervisé est donc délicate. En revanche, des études concordantes ont montré que la convergence était grandement améliorée par la disponibilité d'un petit ensemble de données étiquetées (voir, par exemple, [Nigam et al., 2000] pour la mise en œuvre d'une telle combinaison en catégorisation de documents).

Diverses propositions plus ou moins inspirées d'EM ont également été faites dans un cadre non-probabiliste, émergeant sous le nom *d'apprentissage conjoint* (*co-training*), d'*amorçage* (*bootstrapping*[22]), etc. L'idée générale étant de combiner des sources de connaissance pour transformer des données non-étiquetées en données annotées, lesquelles permettront d'apprendre de meilleurs modèles.

Parmi les principales contributions sur cette question, citons en particulier [Yarowsky, 1995], qui propose une méthodologie originale pour améliorer un système de désambiguïsation sémantique, en s'appuyant sur les heuristiques suivantes :

> The algorithm avoids the need for costly hand tagged training data by exploiting two powerful properties of human language :
> – **One sense per collocation** : nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order, and syntactic relationship
> – **One sense per discourse** : the sense of a target word is highly consistent within any given document.

Partant d'un ensemble réduit de données annotées, l'idée est d'en extraire un ensemble de collocations qui permettront d'étiqueter automatiquement plus de données, qui elles mêmes permettront d'augmenter la liste des collocations, afin, de proche en proche, d'exploiter au maximum les informations contenues dans les données disponibles. La seconde heuristique (*one sense per discourse*) est utilisée pour filtrer des collocations douteuses ou corriger des erreurs produites par la première heuristique.

Dans un contexte d'extraction d'information, [Riloff, 1999, Jones et al., 1999] s'appuient sur une intuition identique : l'observation d'un petit ensemble d'entité nommées de type t permet de révéler des contextes syntaxiques prototypiques pour cette classe ; lesquels contextes permettront de détecter de nouveaux candidats pour ce type, qui à leur tour permettront de détecter de nouveaux contextes...

Une idée proche est finalement mise en œuvre sous le nom de *co-training* dans [Blum and Mitchell, 1998, Denis et al., 2003] pour améliorer un système de catégorisation de documents ; et dans [Collins and Singer, 1999] dans un cadre de la tâche de repérage des entités nommées. Cette approche repose sur l'utilisation *conjointe* de deux systèmes de catégorisation se renforçant mutuellement dans un processus itératif : chaque catégoriseur étiquette des données qui sont ensuite utilisées pour améliorer les performances de l'autre système, qui sera alors en mesure d'étiqueter de nouvelles données, etc. [Abney, 2002] présente une étude comparée de ces différents modèles, en explicitant en particulier les conditions dans lesquelles ils sont efficaces.

[22] J'utilise ici le terme *bootstrapping* avec un sens assez différent du sens traditionnel en statistique, où il fait référence à une technique de rééchantillonage. Ici, *bootstrapping* réfère simplement aux méthodes qui utilisent un faible ensemble de données de supervision pour guider un apprentissage essentiellement non-supervisé.

1.3.2 Le choix d'une représentation

La construction de la représentation des instances d'apprentissage constitue un enjeu central pour aboutir à des catégoriseurs efficaces ou à des regroupements interprétables. Cette question prend un sens particulier en traitement automatique des langues, car le nombre de descripteurs potentiellement mobilisables (et potentiellement utiles) est souvent très grand : d'une part, parce que la fenêtre de contexte autour de l'unité à analyser peut être étendue pratiquement à volonté ; d'autre part, parce que la description linguistique fournit des descripteurs de plusieurs niveaux, qu'il peut être souhaitable d'utiliser conjointement.

Ainsi, pour une tâche de repérage des entités nommées, il peut être utile d'utiliser simultanément des attributs de la forme courante et des formes avoisinantes, ces attributs exprimant des propriétés typographiques et morpho-syntaxiques, voire des informations de nature sémantique. Les modèles exponentiels de phrase [Rosenfeld et al., 2001], introduits pour des tâches de modélisation du langage, intègrent des attributs portant sur la fréquence de paires ou de triplets de mots (pas nécessairement adjacents), sur des configurations syntaxiques, ainsi que des mesures de la cohérence sémantique du vocabulaire.

Si les descripteurs sont potentiellement nombreux, pour une observation donnée, le nombre de descripteurs actifs est souvent faible par rapport au nombre total de descripteurs. Cette situation est illustrée par les tâches de fouille de texte : dans le modèle du sac de mots, chaque document est un point dans un espace de très grande dimension ; toutefois le nombre de composantes non nulles (les mots du documents) est en général très faible au regard de la dimension du modèle (le vocabulaire contenu dans tous les textes). [Roth, 1998] en tire argument pour l'utilisation d'un séparateur linéaire simple bien adapté à cette situation.

La manipulation de ces très gros modèles statistiques pose deux types de problèmes : des problèmes d'estimation et des problèmes computationnels. Ces deux difficultés, ainsi que diverses propositions pour y remédier, sont discutées ci-dessous.

Les distributions d'événements linguistiques

Une problématique commune à tous les apprentis est celle de la qualité des estimations réalisées et de leur capacité de généralisation à de nouvelles données. La théorie statistique de l'apprentissage apporte à cette question des réponses théoriques, exprimées par la vitesse (en nombre d'observations) de convergence des estimateurs. Ces résultats sont en général peu exploitables en traitement automatique des langues, car les modèles statistiques empiriquement les plus performants sont le plus souvent estimés sur des corpus pour lesquels le nombre d'observations est bien inférieur au nombre de paramètres.

Ainsi en modélisation du langage, l'apprentissage d'un modèle 3-gramme sur un vocabulaire de 20 000 mots conduit à estimer de l'ordre de 10^{13} paramètres, alors que les plus gros corpus utilisés peinent à atteindre le milliard de mots[23]. De même, pour des tâches de catégorisation de documents, le corpus 20-Newsgroups ne contient que 20 000 documents pour distinguer les 20 classes : dès lors qu'il utilise un vocabulaire d'indexation de plus de 1 000 formes simples, un catégoriseur bayésien naïf intégrera plus de paramètres qu'il n'y a d'observations (de documents). Enfin, pour reclasser des arbres d'analyse, le modèle utilisé par [Charniak and Johnson, 2005] inclut des millions d'attributs syntaxiques, dont les poids sont estimés sur un ensemble de quelques dizaines de milliers de phrases.

Dans le cadre de tâches du traitement automatique des langues, pour lesquelles les descripteurs associés aux instances à classer concernent typiquement des occurrences de formes ou de lemmes, cette question est aggravée par la forme particulière des distributions d'occurrences. Les travaux originels de Zipf, de Mandelbrot et ceux plus récents de Baayen [2001] ont mis en évidence la forme particulière des distributions des occurrences de mots. Ces distributions se caractérisent par un petit nombre de types très fréquents, alors

[23] La situation est toutefois en train d'évoluer.

que la majorité des types est beaucoup plus rarement observée. Cette observation conduisant Zipf à relier la fréquence f d'un type et son rang r dans un classement par fréquence selon : $f \times r = \text{Cste}$.

Une conséquence en est que la majorité des paramètres des modèles probabilistes doivent être estimés à partir de décomptes très faibles, conduisant à des estimateurs ayant une forte variance. Face à cette difficulté, deux approches (non-exclusives l'une de l'autre) sont possibles : réduire la dimensionalité du problème, par exemple en sélectionnant les descripteurs et mettre en œuvre des techniques de lissage. Quoi qu'il en soit, l'utilisation de techniques de traitement automatique des langues impose de manipuler de très gros modèles, et donc de disposer des outils computationnels adéquats. Ces trois questions sont abordées dans les paragaphes qui suivent.

Sélection d'attributs

La question de la sélection d'attributs a fait l'objet de multiples études dans le domaine de l'apprentissage automatique et de nombreuses stratégies sont envisageables pour effectuer cette sélection. Une des plus simples consiste à repérer les descripteurs les plus pertinents sur la base de critères statistiques (le χ^2) ou informationnels (l'information mutuelle).

Dans le cadre de tâches de traitement automatique des langues, cette stratégie est largement inopérante, dans la mesure où les attributs sont souvent très corrélés entre eux. Évaluer un attribut séparément des autres ne permet pas de prédire correctement sa contribution finale aux performances du modèle. À défaut de pouvoir sélectionner l'ensemble d'attributs permettant d'atteindre les performances optimales, il reste possible de mettre en œuvre des techniques heuristiques, consistant soit à éliminer les attributs trop rares, soit à ajouter (respectivement supprimer) de manière incrémentale les attributs qui améliorent le plus (resp. détériorent le moins) la fonction objectif.

La première approche est couramment utilisée en modélisation du langage, une pratique usuelle consistant à éliminer les k-grammes dont la fréquence d'occurrence est faible, afin de diminuer l'emprise mémoire globale de ces modèles pour des applications de reconnaissance vocale ; de même, en catégorisation de documents, on supprimera du vocabulaire d'indexation les formes dont la fréquence d'occurrence est trop rare. Compte-tenu de la forme des distributions d'occurrence, cette méthode conduit à grandement simplifier les modèles.

Une stratégie plus élaborée est proposée par [Berger et al., 1996] dans le cadre des modèles du maximum d'entropie, qui propose divers algorithmes gloutons pour ajouter itérativement de nouveaux descripteurs dans le modèle, l'utilité marginale d'un descripteur supplémentaire étant évaluée empiriquement sur des données de validation. Des stratégies similaires sont mises en œuvre dans [Rosenfeld et al., 2001] sur des modèles exponentiels de phrase, ou encore dans [McCallum, 2003] qui l'adapte au cas des champs aléatoires conditionnels.

Il apparaît au final que le principal bénéfice de l'utilisation de ces techniques est de diminuer le coût computationnel de l'utilisation de gros modèles. En revanche, si seules comptent les performances, il semble que la meilleure stratégie consiste à conserver un maximum des informations du corpus d'apprentissage, en cherchant à limiter les problèmes d'estimation par utilisation de techniques de lissage.

Lissage

La problématique du lissage surgit de manière récurrente en apprentissage automatique ; en traitement automatique des langues, elle a surtout fait l'objet de travaux dans le cadre de l'estimation de modélisation du langage : l'immense majorité des paramètres d'un modèle k-gramme modélisant des événements qui ne sont jamais observés lors de l'apprentissage. Comme le montrent les résultats empiriques de

[Chen and Goodman, 1996], le choix d'une technique de lissage conditionne fortement les performances des modèles de langage.

Une réflexion similaire a également été conduite dans le contexte de l'estimation de modèles du maximum d'entropie par [Chen and Rosenfeld, 2000, Goodman, 2004], qui proposent d'introduire des distributions gaussiennes sur les paramètres du modèle. Dans le cadre de la catégorisation de documents, l'estimation de modèles bayésiens naïfs implique également un lissage des distributions conditionnelles. Ainsi, lorsque les vecteurs de comptes de mots sont modélisés par des distributions multinomiales [McCallum and Nigam, 1998], ce lissage est opéré en posant une distribution *a priori* de Dirichlet sur les paramètres des lois multinomiales.

Ces techniques semblent, pour les tâches supervisées, suffisantes pour se prémunir effectivement contre les problèmes de sur-apprentissage. Dans ce contexte, il semble meilleur de conserver un maximum d'information, quitte à avoir des estimateurs peu robustes, plutôt que chercher à filtrer les attributs, ce qui induit le plus souvent une dégradation des performances. La question se pose sans doute différemment en ce qui concerne les tâches d'apprentissage non-supervisé (les travaux présentés à la section 3.3 dans un cadre de classification de textes confirment cette observation).

Quelques problèmes computationnels

Accepter de travailler en très grande dimension pose des problèmes computationnels difficiles : les modèles de langage utilisés en reconnaissance vocale demandent le stockage en mémoire de millions de paramètres ; lorsque le décodage s'appuie sur une expansion complète de l'espace de recherche, il est courant de manipuler des graphes contenant des dizaines de millions de transitions, suscitant de nombreux travaux pour optimiser la forme de ce graphe ainsi que les procédures de stockage et de parcours associées (voir, sur ces questions, la section 4.2).

Pour des modèles plus complexes que les k-grammes, les difficultés se posent dès l'apprentissage : de nombreux modèles statistiques expriment l'apprentissage comme un problème d'optimisation en très grande dimension. C'est en particulier le cas des modèles de type maximum d'entropie (ou CRFs) qui, lorsqu'on les estime par des techniques d'optimisation numériques classiques (descente de gradient, méthode de Newton ou quasi Newton), demandent à chaque itération d'évaluer des sommes de produits scalaires en très grande dimension. En ce qui concerne les modèles de transcodage, la restriction à des dépendances locales entre symboles de la sortie permet d'utiliser pour ces calculs des algorithmes efficaces (programmation dynamique, algorithme *forward-backward*).

Par comparaison, les méthodes à noyaux [Vapnik, 1995, Scholkopf and Smola, 2001] bénéficient d'un avantage certain, puisqu'elles se fondent sur l'évaluation du produit scalaire entre les représentations des instances, qui sont typiquement très creuses. De nombreux travaux récents proposent des noyaux computationnellement efficaces pour des représentations séquentielles ou hiérarchiques (voir, par exemple, [Cortes et al., 2002, Collins and Duffy, 2002, Vishwanathan and Smola, 2003]), utilisant en particulier des techniques à base de transducteurs finis. Cette capacité à s'accomoder de descriptions évoluant dans des espaces de très grande dimension [Joachims, 1999] et incluant des dépendances entre descripteurs a largement contribué à la popularité des SVMs pour des applications de traitement automatique des langues.

1.3.3 Estimation des modèles

De nombreuses tâches du traitement automatique des langues se posent naturellement comme des problèmes de désambiguïsation de la meilleure structure à assigner à une nouvelle entrée. D'un point de vue formel, ils s'expriment donc comme des problèmes de discrimination. Deux types d'approches sont alors possibles, consistant à utiliser des modèles génératifs ou des modèles discriminants (c.f. la section 1.1.2).

Modèles discriminants vs. génératifs Pour les problèmes de discrimination entre un nombre fini de catégories, l'expérience, tirée de multiples évaluations comparatives, plaide en faveur de l' utilisation de modèles discriminants.

Il n'est pas surprenant dès lors que la conception et l'évaluation de modèles discriminants pour les séquences ait suscité un effort considérable de la communauté, donnant lieu, pour les tâches de segmentation et de balisage, à de multiples propositions pour trouver des équivalents discriminants aux modèles (génératifs) de Markov cachés. Cet effort a aboutit au développement de modèles effectivement plus performants que les HMMs, tels que les champs aléatoires conditionnels [Lafferty et al., 2001] (voir également [Altun et al., 2002, 2003]). Ces modèles conservent toutefois certaines des hypothèses principales des HMMs, en particulier la localité des dépendances conditionnelles entre valeurs de la séquence de sortie (y). Cette hypothèse permet en effet d'utiliser des techniques de programmation dynamique pour l'estimation comme pour le décodage.

La même évolution est observée en analyse syntaxique probabiliste : alors que le parseur stochastique de [Charniak, 1993] est estimé au maximum de vraisemblance, [Johnson, 2001, Rozenknop, 2002] proposent des algorithmes pour l'estimation discriminante de grammaires hors-contexte stochastiques.

La question de la supériorité absolue d'une approche sur l'autre reste toutefois controversée : [Johnson, 2001] présente des comparaisons contrastées entre les deux types d'approche, que les arguments théoriques et empiriques de [Ng and Jordan, 2002] permettent partiellement d'expliquer : les modèles discriminants sont asymptotiquement plus performants, mais demandent plus de données d'apprentissage que les modèles génératifs. Dans le cas de tâches d'apprentissage relatives à des données linguistiques, la rareté relative des données conduit à mieux estimer les modèles génératifs.

Les résultats de [Klein and Manning, 2002] tendent à nuancer ce jugement : même en présence de faibles quantités de données, optimiser un critère discriminant peut être meilleur que maximiser la vraisemblance d'un modèle génératif. Comme le concluent ces auteurs, l'avantage des modèles discriminants semble s'expliquer par le fait que la modélisation de la loi (conditionnelle) des observations n'est pas nécessaire, ce qui permet de prendre en compte des descripteurs arbitraires, y compris très inter-dépendants.

Apprentissage discriminant Faute de pouvoir utiliser un modèle discriminant, il est possible d'optimiser, dans un cadre supervisé, les paramètres d'un modèle génératif pour minimiser une fonction du taux d'erreur (plutôt que la vraisemblance des observations). La supériorité de critères discriminants sur des critères tels que le maximum de vraisemblance a été démontrée dans le contexte de l'estimation de modèles acoustiques [Valtchev et al., 1997, Schlüter et al., 2001] et linguistiques [Chen and Rosenfeld, 2000, Kuo et al., 2002] pour la reconnaissance vocale (voir également la section 4.2.5 pour notre contribution sur ce thème).

Une approche voisine et relativement générique a déjà été mentionnée à la section 1.2.4, dans laquelle nous avons présenté le cadre du réordonnancement (*reranking*), originellement introduit dans [Collins and Duffy, 2002].

1.3.4 Quelques perspectives

Les discussions des paragraphes précédents ont conduit à identifier un certain nombre de questions générales qui me semblent encore mal résolues dans le cadre de l'apprentissage statistique. Pour achever cette discussion, j'ai choisi de développer deux thèmes, pour lesquels divers développements des travaux en cours sont attendus : la prise en compte des dépendances et l'utilisation de très grosses bases de données.

Avant de débuter cette présentation, notons qu'il existe naturellement de nombreux autres sujets importants en apprentissage statistique. Parmi ceux qui me semblent pertinents pour le traitement automatique des langues, mais qui ne seront pas développés ici, mentionnons en particulier les questions de sélection de

modèle, ainsi que les problématiques liées à une meilleure utilisation des ressources linguistiques.

Plus d'inter-dépendances

La modélisation des dépendances entre les décisions dans les modèles utilisés pour la catégorisation, la modélisation de séquences, ou le transcodage, continue d'être un enjeu important. Ces dépendances restent en effet imparfaitement modélisées : pour les modèles qui travaillent sur des séquences dont l'unité est la phrase, seules des dépendances locales et internes à la phrase sont effectivement prises en compte, alors même que (i) il existe des dépendances à longue distance au sein des phrases et (ii) il existe des dépendances riches entre phrases d'un même document, que l'ingénieux apprenti proposé par [Yarowsky, 1995], qui tire parti de l'information que deux occurrences de la même forme apparaissent au sein d'un même document, a clairement mis en évidence.

Une meilleure modélisation de ces dépendances est donc susceptible d'avoir un impact positif sur les performances de nombreux systèmes[24]. C'est ce que montre, pour des tâches de modélisation statistique du langage, les travaux de Rosenfeld [1994], qui cherche à capturer l'influence de mots situés « à longue distance » du mot courant. L'exploitation de ces contextes élargis reste toutefois algorithmiquement difficile, les difficultés se posant à la fois pour identifier les dépendances utiles et pour les exploiter efficacement au sein des systèmes de reconnaissance vocale.

Une autre tâche pour laquelle ces dépendances gagneraient à être mieux modélisées est le repérage des entités nommées : les entités sont identifiées au sein d'un texte de manière indépendante. Pourtant, lorsqu'une entité est ambiguë, il est clair qu'il y aurait à gagner à exploiter tous les contextes disponibles et affecter la même étiquette pour toutes les occurrences au sein d'un document. Une première tentative dans cette direction consiste à utiliser un premier balisage automatique pour ré-entraîner le système d'étiquetage de manière à le rendre plus à même de prendre des décisions cohérentes [Mikheev et al., 1999]. Il apparaît également que la reconnaissance d'une entité rend beaucoup plus probables certaines entités, et en interdit d'autres. Ainsi, par exemple, pour toutes les tâches d'extraction d'information orientées vers le remplissage d'un formulaire, l'identification d'un fragment référent par exemple à la date de l'événement devrait rendre impossible l'identification d'une seconde date. Là encore, des travaux commencent à exister, qui tentent de modéliser la tâche d'annotation d'un document comme un problème d'optimisation global, plutôt que local [Finkel et al., 2005].

Pour donner un troisième exemple, les systèmes de traduction statistique effectuent leur choix de traduction phrase par phrase. Il est dès lors possible que diverses occurrences d'un mot donné au sein d'un texte soient traduites de plusieurs manières. Il est patent qu'une vision plus globale des décisions prises ne peut que conduire à une amélioration de leur cohérence.

La question de la modélisation des dépendances longues au sein des flux d'entrées ou de sorties, ou encore de contraintes structurelles non-locales dans les sorties est un thème de recherche actif[25], au sujet duquel beaucoup reste encore à accomplir. Cette question est non triviale à résoudre pour les modèles de séquence, car la mise en œuvre des transductions statistiques repose au fond sur un socle algorithmique très efficace (principalement la programmation dynamique, instancié par les algorithmes de Viterbi et de Baum-Welsh pour les HMMs ou leurs équivalents dans les autres modèles) mais qui impose des contraintes fortes de dépendances localisées. S'en affranchir imposera donc de renoncer à des techniques d'estimation et de décodage à la fois bien connues et efficaces, au profit d'autres techniques d'estimation (par exemple des techniques de simulation comme les *Monte Carlo Markov Chains*) et d'optimisation combinatoire aujourd'hui plus coûteuses à

[24] Cet impact est sans doute très dépendant des cadres applicatifs. Le contre-exemple est celui de la recherche d'information, dans lesquels les performances du modèle du « sac de mots » restent difficiles à surpasser.

[25] On consultera, par exemple, [Taskar et al., 2003, Tsochantaridis et al., 2004, Bunescu and Mooney, 2004] pour un échantillon de travaux récents.

mettre en œuvre et moins bien maîtrisées.

Encore plus de données

Nous l'avons évoqué dans la section 1.3.1, les modèles couramment utilisés en traitement automatique des langues sont estimés sur des ensembles de données trop réduits pour qu'on puisse aboutir à des estimateurs fiables et donc à des performances optimales. Le travail expérimental de [Banko and Brill, 2001] montre, sur un problème jouet, le gain qu'il est possible d'obtenir en exploitant des corpus beaucoup plus volumineux que ceux utilisés aujourd'hui.

Une telle évolution est certainement susceptible d'apporter, dans le court terme, des améliorations sensibles des performances des systèmes statistiques[26]. Ne manque pour cela que la disponibilité de « méga-corpus » suffisamment propres, annotés et bien documentés, l'utilisation directe des ressources disponibles sur l'Internet ayant été une relative source de désillusion [Kilgariff and Grefenstette, 2003], du fait de (i) la quasi-obligation d'utiliser les portails de recherche généralistes, dont l'accès est de plus en plus restreint et difficile et (ii) l'immense difficulté à traiter les collections collectées sur le Web, qui posent des problèmes d'ingénierie (gestion des formats, filtrage du spam) redoutables.

Cette évolution pose dès aujourd'hui des problèmes computationnels réels : si le passage à l'échelle a déjà été opéré pour les systèmes de recherche d'information, il est loin d'être fait pour les différents modules impliqués dans les tâches de traitement automatique des langues, et les composants qui reposent sur des techniques d'apprentissage n'échappent pas à cette critique. Un important travail algorithmique est donc nécessaire et attendu autour des problématiques de la fouille de très gros volumes de données textuelles, du développement d'algorithmes parallélisables et/ou exploitant des grilles de calcul. Des travaux dans cette direction commencent à émerger au sein de la communauté ainsi, par exemple, [Pantel et al., 2004]. La récente initiative de Google, qui annonce la distribution de statistiques accumulées sur des volumes de données incomparables[27], va certainement contribuer à amplifier les efforts dans cette voie.

À plus long terme, il n'est toutefois pas acquis que l'accumulation de plus de données, pour autant qu'on sache les exploiter, sera une solution adaptée à tous les problèmes. Compte-tenu des distributions d'occurrence des événements de la langue et de la tendance envers l'utilisation de modèles plus complexes, une autre direction de recherche, également prometteuse, consiste à développer des algorithmes d'apprentissage actif et d'apprentissage en ligne. Rechercher sélectivement de nouveaux exemples difficiles à traiter, mettre en place des stratégies de mise à jour incrémentales et continues des modèles sont autant de moyens d'optimiser les capacités de généralisation et maintenir des modèles qui s'adaptent aux évolutions des usages de la langue. On ne peut que s'étonner du faible nombre de travaux s'intéressant à ces problématiques, qui sont pourtant bien connues en apprentissage automatique, dans un cadre de traitement automatique des langues.

1.4 Un bilan contrasté

Au terme de ce large tour d'horizon, il nous semble possible de dresser un premier bilan, doublement contrasté de ces quelques quinze années de recherche.

Bilan contrasté, dans la mesure où loin de conduire au développement d'apprentis et de biais bien adaptés à la forme et aux spécificités des données linguistiques[28], cette rencontre a, nous l'avons vu, majoritaire-

[26] Voir toutefois les réserves que je pose au début de la section 1.3.1.

[27] cf. `http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html` : « We processed 1,011,582,453,213 words of running text and are publishing the counts for all 1,146,580,664 five-word sequences that appear at least 40 times. There are 13,653,070 unique words, after discarding words that appear less than 200 times. »

[28] Voir, toutefois, les travaux déjà cités de D. Klein et C. Manning [2002] ; ceux de D. Roth [1998, 1999, 2000] et de W. Daelemans

ment suscité des travaux visant à mettre en œuvre, au prix parfois de pré-traitements hardis, des méthodes généralistes de l'apprentissage automatique.

Avant d'entamer cette discussion, notons que certains des arguments avancés dans cette section seront développés dans le chapitre 2, consacré à mes propres travaux sur *l'apprentissage par analogie* : sans anticiper sur cette présentation, mentionnons qu'une des ambitions sera précisément de discuter pourquoi l'analogie, par opposition par exemple avec des apprentis statistiques, me semble être un bon exemple de mécanisme d'inférence particulièrement adapté à certains types de traitements linguistiques.

1.4.1 Des évolutions significatives

Le renouveau des méthodes empiriques en traitement automatique des langues a eu des effets très sensibles en traitement automatique des langues : il a contribué à changer la manière d'envisager les traitements linguistiques sous la forme de *modules* plus ou moins indépendants ; et à accorder une importance croissante à *l'évaluation des performances* des systèmes. Les conséquences de ces deux évolutions sur le domaine sont toutefois ambivalentes.

Au rebours, la fréquentation des données linguistiques, qui présentent des spécificités très marquées, n'a, me semble-t-il, que très marginalement contribué aux évolutions des problématiques de l'apprentissage automatique. L'évolution la plus sensible est, de mon point de vue, l'obligation d'un changement d'échelle qu'implique le traitement de telles données.

L'impression générale qui se dégage toutefois de cette rétrospective est donc une emprise croissante des préoccupations pragmatiques qui incitent à utiliser des techniques d'apprentissage simples et robustes, lesquelles garantissent souvent, à bon compte, des performances satisfaisantes.

Le traitement automatique des langues transformé

Modularisation La multiplication des travaux exploitant des outils d'apprentissage automatique s'est accompagnée d'une « modularisation » progressive du domaine, permettant d'identifier des composants standard, des entrées-sorties, des sous-tâches évaluables (désambiguïsation sémantique, repérage des entités nommées, alignement de corpus, etc.). Si cette modularisation est souvent bienvenue et permet au concepteur de système de combiner librement les composants qu'il souhaite intégrer, il me semble toutefois également nécessaire de s'interroger sur les effets pas toujours positifs de cette évolution.

Du point de vue de la modélisation, cette modularisation a conduit à la spécification de sous-tâches dont la pertinence linguistique me paraît contestable : ainsi par exemple, [Boyd et al., 2005] traitent comme un « pur » problème de catégorisation supervisée la question de la discrimination des emplois référentiels et non-référentiels[29] du pronom troisième personne '*it*'. Cette question est posée comme une sous-tâche préalable au traitement des anaphores pronominales, qui est elle même une sous-partie du problème de traitement des co-références. Il n'est pas clair que la spécification de cette tâche, l'identification des descripteurs qui permettent de la réaliser avec succès, ait beaucoup fait progresser la compréhension du problème.

---

et A. van den Bosch [Daelemans and van den Bosch, 2005]

[29] '*it* est référentiel lorsqu'il est utilisé en lieu et place d'un groupe nominal auquel il renvoie, comme dans

(1) *The car broke ; it was immediately replaced* ;

à l'inverse, les emplois non-référentiels correspondent à des formes impersonnelles, ou encore à des idiomes, et ne donnent pas lieu à la recherche de l'antécédent, comme dans

(2) *We had a terrible weather : it rained all day long*

.

Même si l'on se place du point de vue des performances des systèmes, il n'est pas évident que cette modularisation ait été entièrement bénéfique : la multiplication des tâches intermédiaires complique singulièrement la forme du *pipe-line* le long duquel s'enchaînent les traitements nécessaires à une tâche finale. Chaque sous-tâche risque toutefois d'entraîner des erreurs, qui se propagent et se cumulent le long de la chaîne de traitement, sans que l'on sache vraiment évaluer le rythme de cette accumulation, ni mesurer l'impact des erreurs d'un module sur les modules qui lui succèdent.

Autre effet négatif, cette modularisation a, me semble-t-il, également entraîné un cloisonnement des différentes communautés travaillant autour de « petites » tâches, conduisant à masquer la grande similarité des outils qu'elles employaient, similarité que nous avons essayé de mettre en évidence dans les sections précédentes.

Le culte de la performance La diffusion des méthodes statistiques a également fortement orienté la communauté vers la recherche de systèmes atteignant des performances optimales pour les métriques de la tâche. En témoignent la multiplication des « compétitions », challenges et autres défis qui se multiplient en marge des principales conférences du domaine. Cet intérêt grandissant pour les questions de l'évaluation des performances des outils de traitement automatique des langues est certainement une évolution positive, qui permet d'apprécier de manière objective les progrès réalisés, et suscite les échanges sur des problématiques et des données communes.

On notera qu'à ce compte, pour de multiples tâches, les outils industriels les plus performants restent ceux qui s'appuient sur des outils symboliques, exploitant des bases de connaissance d'experts accumulées au fil des ans. Pour prendre un exemple, s'il est relativement simple de bâtir un étiquetteur morpho-syntaxique statistique du français démontrant des performances très acceptables, il est probablement bien plus difficile de surpasser les performances du meilleur étiqueteur à base de règles. La même constatation pourrait être faite sur d'autres tâches, comme la conversion graphème-phonème, l'extraction d'information ou encore la traduction automatique.

Sur la question de l'évaluation toujours, on déplorera néanmoins que l'évaluation quantitative des modules de traitements, réalisée le plus souvent de manière indépendante des autres modules, soit devenu le « mètre-étalon » unique des apprentis linguistiques. Les autres évaluations possibles des travaux liant apprentissage et traitement automatique des langues : leur adéquation avec une théorie linguistique ou avec des observations comportementales, leur apprenabilité, leur plausibilité cognitive, leur stabilité (qu'on pense à des tâches non-supervisées), l'interprétabilité des modèles induits ou de leurs paramètres, sont ainsi progressivement passés au second plan des préoccupations.

Apprentissage, évolution du langage Paradoxe apparent, la multiplication des travaux ayant recours à des techniques d'apprentissage n'a que très marginalement contribué à la compréhension des mécanismes impliqués dans l'acquisition et dans l'évolution naturelle du langage. Cette thématique, pourtant très bien représentée dans les premiers temps de cette réhabilitation des méthodes empiriques en traitement automatique des langues, s'est progressivement estompée et n'apparaît plus que marginalement dans les principales conférences du domaine[30]. La rareté des corpus naturels (en dépit d'initiatives telles que CHILDES[31]) est naturellement un frein important au développement de ces recherches, mais on ne peut manquer de faire le lien entre le passage au second plan de ces problématiques plus linguistiques ou psycho-linguistiques et l'augmentation des efforts consentis pour concevoir des apprentis performants sur des couches intermédiaires de traitement de plus en plus spécifiques.

[30] Elle reste toutefois bien représentée dans les conférences de sciences cognitives, mais l'interpénétration des deux communautés reste relativement faible.

[31] http://childes.psy.cmu.edu/

Des apprentis plus puissants

Le développement d'apprentis performants pour des sous-tâches bien identifiées a conduit à une sophistication des méthodes d'apprentissage utilisées en traitement automatique des langues, avec des innovations dans la modélisation des dépendances locales entre éléments adjacents d'une séquence, ainsi que dans la mise en place de stratégies d'apprentissage discriminant. Pour l'essentiel, cette évolution reste marquée par une nette prédominance des techniques d'apprentissage statistique, exploitant des représentations *simples* (attribut-valeur), dans des espaces de *grande dimension* et des régimes d'apprentissage *standardisés*.

Débutons par le troisième terme de cette affirmation : l'apprentissage automatique connaît divers régimes d'apprentissage : en ligne (incrémental) ou en batch ; supervisé ou non-supervisé, semi-supervisé ou encore par renforcement ; avec ou sans interaction avec un humain. Chacun de ces régimes d'apprentissage modélise un cadre applicatif ou une situation d'apprentissage naturel particulier, en spécifiant notamment les interactions avec la source des données et avec le tuteur (qui fournit les données de supervision). On l'a vu, les applications du traitement automatique des langues restent toutefois marquées par une prédominance des méthodes d'apprentissage supervisé en batch, sans doute peut-être les plus en phase avec l'obligation d'une mesure qualitative des performances.

Avant de commenter les deux autres termes, il me semble important de revenir sur certaines spécificités des données linguistiques.

Spécificités des données linguistiques Comme j'ai essayé de le souligner tout au long de ce chapitre, les données linguistiques présentent de nombreuses particularités qui les distinguent d'autres types de données. Rappelons, en particulier, que :

(i) ces données sont le plus souvent symboliques, fortement structurées, relationnelles, hiérarchiques et multi-dimensionnelles ;

(ii) elles présentent ensuite la particularité d'avoir été l'objet de nombreuses études et le savoir accumulé sur ces données (au moins pour certaines langues) est abondant. On sait en particulier l'extrême difficulté de construire un modèle pour décrire certaines strates de la connaissance linguistique.

(iii) elles sont aujourd'hui disponibles dans des proportions qui excèdent de loin les données disponibles pour les problèmes traditionnellement considérés par les apprentistes [Hettich et al., 1998].

(iv) la distribution statistique des unités linguistiques présente, à quelque niveau d'observation qu'on se situe, des caractéristiques tout à fait particulières : un très petit nombre de formes (ou règles) suffit à expliquer la majorité des occurrences ; un très grand nombre de formes (ou de règles) est nécessaire pour expliquer la totalité des occurrences.

Apprentissage symbolique Il me semble que les deux premières particularités ont pesé pour peu dans les évolutions du domaine. En effet, les apprentis symboliques exploitant des données relationnelles et/ou des biais linguistiques ((i) et (ii)), en particulier les méthodes de programmation logique inductive (ILP), s'ils ont pu rencontrer quelques succès [Zelle and Mooney, 1993, Califf and Mooney, 1995, 1997], sont demeurés relativement marginaux en traitement automatique des langues. Il en va de même des techniques d'inférence grammaticale [Dupont and Miclet, 1998], qui, bien que spécifiquement adaptées au traitement des données séquentielles ou plus généralement aux données structurées (graphes, arbres), restent en pratique peu utilisées dans un cadre de traitement automatique des langues (voir toutefois, par exemple, [Gildea and Jurafsky, 1994, Oncina et al., 1994]).

Apprentissage paresseux À la différence des modèles probabilistes évoqués dans ce chapitre, les apprentis « paresseux » [Aha et al., 1991, Mitchell, 1997] généralisent directement à partir l'ensemble d'apprentissage, s'abstenant de construire une abstraction (un modèle) de ces données. Ils évitent ainsi un des

problèmes des approches à base de règles, qui peinent à prendre en compte des idiosyncrasies, exceptions et micro-régularités (point (iv)). Remplacer les données par un modèle semble conduire à perdre une information potentiellement utile [Daelemans et al., 1999]. Le peu d'écho que cette observation a eue, hormis en traduction automatique où les apprentis paresseux restent populaires [Carl and Way, 2003], reste une source d'étonnement, d'autant plus que ces systèmes permettent souvent d'atteindre des performances proches de l'état de l'art. En présentant nos propres travaux sur les modèles à base d'analogie au chapitre 2, nous aurons l'occasion de revenir sur les bénéfices d'une telle stratégie d'apprentissage.

Apprentissage statistique Plus surprenant encore peut-être, alors que l'emprise des méthodes statistiques se développait, la réalisation de la non-normalité des distributions d'événements linguistiques a suscité bien peu de tentatives de modéliser plus fidèlement ces décomptes ; ou encore pour tirer les leçons de l'existence de phénomènes idiosyncrasiques.

Pour l'essentiel, les principales réalisations ont donc tiré profit de la capacité qu'ont certains apprentis statistiques simples (exemplairement : des systèmes supervisés de catégorisation construisant des *séparations linéaires*) à exploiter efficacement de très gros volumes de données, dans des espaces de grande dimension. Cette approche présente un double intérêt. En premier lieu, la disponibilité de corpus d'apprentissage de grande taille dispense souvent de devoir exploiter les éventuelles connaissances linguistiques qui seraient disponibles. En second lieu, l'utilisation de transformations idoines, par lesquelles les données structurées sont converties en des vecteurs attribut-valeur numériques de grande dimension, évite de devoir modéliser des dépendances complexes et/ou d'avoir à manipuler des représentations structurées.

1.4.2 Un mariage de convenance ?

En résumé, l'impression qui se dégage de cette rétrospective est assez conforme à ce que l'on peut observer plus généralement dans le domaine du traitement automatique des langues, et que l'on peut résumer brutalement par une place croissante prise par *l'ingénierie linguistique* au détriment de *l'informatique linguistique*. En témoignent, nous l'avons noté, la quasi disparition dans les conférences du domaine des travaux consacrés à l'apprenabilité de modèles linguistiques particuliers, ou bien encore visant à élaborer des modèles de l'apprentissage humain.

Cette évolution a logiquement accompagné la progressive numérisation des traditionnels véhicules des énoncés linguistiques : numérisation de la voix (téléphone), de l'image fixe et animée (télévision, cinéma), ainsi que des principaux supports des textes (mails, Internet, collections numérisées), numérisation qui suscite une demande sociale croissante pour des outils de traitement performants.

Les bailleurs publics n'ont pas manqué de relayer cette demande, en demandant aux laboratoires spécialisés en traitement automatique des langues qu'ils proposent des outils capables de résoudre des tâches bien définies : classer, résumer, rechercher, traduire des documents, extraire des informations, etc. ; et en mettant en place des campagnes d'évaluation, incitant les équipes à comparer objectivement les performances de leurs outils sur la base de données et de protocoles de tests publics.

Face à ces demandes de plus en plus fortes, s'exprimant à la fois pour des tâches de production et de gestion de textes numérisés, les outils issus de l'apprentissage statistique se sont souvent avérés immédiatement mobilisables, quand ils n'étaient pas, de surcroît, les plus robustes et partant, les plus performants. La capacité des modèles statistiques à s'appliquer pour un grand nombre de langues et de domaines a été, dans ce contexte, un argument supplémentaire[32] pour que ces outils se diffusent très largement.

[32] Argument utilisé de manière parfois spécieuse : ses promoteurs soulignent volontiers la difficulté à réutiliser les systèmes à base de connaissance à de nouveaux contextes, en omettant de préciser que pour s'adapter, les techniques d'apprentissage nécessitent des données étiquetées, dont la construction a également un coût non-négligeable.

Cette rencontre entre traitement automatique des langues et apprentissage me semble avoir surtout débouché sur des avancées technologiques, qui ont permis aux deux communautés concernées de mettre en valeur leurs savoir-faire et diffuser leurs innovations, au prix sans doute d'une focalisation des thématiques de recherche autour de quelques thèmes très porteurs et d'une « standardisation » parfois regrettable des travaux publiés.

Bibliographie

Steven Abney. Bootstrapping. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 360–367, Philadelphia, PS, 2002.

Steven Abney. Statistical methods and linguistics. In Judith L. Klavans and Philippe Resnik, editors, *The Balancing Act*, pages 1–27, Cambridge, MA, 1996. The MIT Press.

David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6 :37–66, 1991.

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary : A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1 :113–141, 2000.

Yasamin Altun, Thomas Hofmann, and Mark Johnson. Discriminative learning for label sequences via boosting. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS*15)*, Vancouver, Canada, 2002.

Yasamin Altun, Ioannis Tsochandaridis, and Thomas Hofmann. Hidden markov support vector machines. In *Proceedings of the International Conference on Machine Learning*, Washington DC, USA, 2003.

Harald Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 2001.

Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse, France, 2001.

Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34, 1999.

Jean-Paul Benzécri and ali. *Pratique de l'analyse des données, tome3. Linguistique et lexicologie*. Dunod, Paris, 1981.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1) :39–72, 1996.

Douglas Biber. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(219–241), 1993.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA, 2002. MIT Press.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proccedings of 11th annual conference on Computational Learning Theory (COLT 98)*, pages 92–100, Madison, WI, 1998.

Rens Bod. *Enriching Linguistics with Statistics : Performance Models of Natural Language*. Institute for Logic, Language and Computation, Amsterdam, 1995.

Adriane Boyd, Whitney Gregg-Harrison, and Donna Byron. Identifying non-referential it : A machine learning approach incorporating linguistically motivated patterns. *Traitement automatique des langues (TAL)*, 46(1), 2005.

Eric Brill. Automatic grammar induction and parsing free text : a transformation-based approach. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 259–265, Columbus, Ohio, 1993.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Journal of Natural Language Engineering*, 16(2) :79–85, 1990.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4) :467–479, 1992.

Razvan Bunescu and Raymond J. Mooney. Collective information extraction with relational markov networks. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pages 439–446, Barcelona, Spain, 2004.

Mary Elaine Califf and Raymond J. Mooney. Using inductive logic programming to learn the past tense of English verbs. In Stefan Wermter, editor, *Proceeding of the IJCAI'95 workshop on 'New Appoaches to Learning for Natural Language Processing'*, pages 87–94, Montréal, Canada, 1995.

Mary Elaine Califf and Raymond J. Mooney. Applying ILP-based techniques to natural language information extraction : an experiment in relational learning. In *Working Notes of the IJCAI-97 Workshop on Frontiers of Inductive Logic Processing*, 1997.

Michael Carl and Andy Way, editors. *Recent Advances in Example-Based Machine Translation*, volume 21 of *Text, Speech and Language Technology*. Springer, Verlag, 2003.

Eugene Charniak. *Statistical Language Learning*. The MIT Press, Cambridge, MA, 1993.

Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Association for Computational Linguistics*, pages 173–180, Ann Arbor, MI, 2005.

Stanley F. Chen and Joshua T. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, 1996.

Stanley F. Chen and Ronald Rosenfeld. A survey of smoothing techniques for maximum entropy models. *IEEE transactions on Speech and Audio Processing*, 8(2) :37–50, 2000.

Freddy Y. Y. Choi. Advances in domain independant linear text segmentation. In *Proceedings of the Conference of North American Chapter of the ACL*, Seattle, WA, 2000.

Kenneth W. Church. Speech and language processing : Where have we been and where are we going. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, 2003.

Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied NLP*, pages 136–143, Austin, Texas, 1988.

Kenneth W. Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1) :20–28, 1991.

Kenneth W. Church and Robert L. Mercer. Introduction to computational linguistics special issue on large corpora. *Computational Linguistics*, 1(19) :1–24, 1993.

Philip R. Clarkson. *Adaptation of Statistical Language Models for Automatic Speech Recognition*. PhD thesis, School of Computer Science, Cambridge University, 1999.

Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

Michael Collins and Nigel Duffy. New ranking algorithms for parsing and tagging : kernels over discrete structures and the voted perceptron. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 489–496, Philadelphia, PA, 2002.

Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Maryland, 1999. http ://www.research.att.com/ mcollins/.

Marcel Cori and Jacqueline Léon. La constitution du tal. *Traitement automatique des langues (TAL)*, 43 (3) :21–55, 2002.

Antoine Cornuéjols and Laurent Miclet. *Apprentissage artificiel, concepts et algorithmes*. Eyrolle, Paris, 2002.

Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational kernels. In *Proceedings of NIPS 2002*, Vancouver, 2002.

Thomas M. Cover and Joy A. Thomas. *Elememts of Infomation Theory*. John Wiley and Sons, 1991.

Hamish Cunningham. A definition and short history of language engineering. *Journal of Natural Language Engineering*, 5(1) :1–16, 1999.

Walter Daelemans and Antal van den Bosch. *Memory-Based Language Processing*. Cambridge University Press, 2005.

Walter Daelemans, Steven Gillis, Gert Durieux, and Antal van den Bosch. Learnability and markedness in data-driven acquisition of stress. Technical Report 43, ITK, Tilburg University, NL, May 1993.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. MBT : A memory-based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 14–27, Copenhagen, Denmark, 1996.

Walter Daelemans, Antal Van Den Bosch, and Jakub Zavrel. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34 :11, 1999.

Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université Paris VII, Paris, France, février 1994.

Robert Dale, Hermann Moisl, and Harold Somers, editors. *Handbook of Natural Language Processing*. Marcel Dekker, 2000.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26 :101–126, 2006.

Carl de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Dpt of Computer Science, MIT, 1996.

Renato DeMori and Marcello Federico. Language model adaptation. In Keith Ponting, editor, *Computational Models of Speech Pattern Processing*, pages 280–303. Springer Verlag, 1998.

François Denis, Rémi Gilleron, Anne Laurent, and Marc Tommasi. Text classification and co-training from positive and unlabeled examples. In *Proceedings of the ICML Workshop : the Continuum from Labeled Data to Unlabeled Data in Machine Learning and Data Mining*, pages 80–87, Washington DC, 2003.

Steven J. DeRose. Grammatical category desambiguation by statistical optimization. *Computational Linguistics*, 14(1) :31–39, 1988.

Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2 :263–286, 1995.

Richard 0. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, 2nd edition edition, 2001.

Pierre Dupont and Laurent Miclet. Inférence grammaticale régulière : fondements théoriques et principaux algorithmes. Technical Report 1189, IRISA, Rennes, 1998.

Steve Finch and Nick Chater. Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, pages 820–825, 1992.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pages 148–146. Morgan Kaufmann, 1996.

Catherine Fuchs. *Les ambiguïtés du français*. Ophrys, Paris, France, 1996.

William Gale, Kenneth W. Church, and David Yarowski. Method for disambiguating word senses in a large corpus. *Computer and the Humanities*, 26 :415–439, 1992.

Eric Gaussier. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Methods in Natural Language Processing*, College Park, MD, 1999.

Dan Gildea and Dan Jurafsky. Automatic induction of finite state transducers for simple phonological rules. Technical Report TR-94-052, International Computer Science Institute, Berkeley, CA, 1994.

John Goldsmith. Unsupervised learning of the morphology of natural languages. *Computational Linguistics*, 27(2) :153–198, 2001.

Joshua T. Goodman. A bit of progress in language modeling. Technical Report MSR-TR-2001–72, Microsoft Research, Redmond, WA, 2001.

Joshua T. Goodman. Exponential priors for maximum entropy models. In *Proceedings of the annual meeting of the North American Chapter of the ACL*, Boston, MA, 2004.

Benoît Habert, Adeline Nazarenko, and André Salem. *Les linguistiques de corpus*. U Linguistique. Armand Colin/Masson, Paris, 1997.

Marti Hearst. TextTiling : Segmenting texts into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1) :33–64, 1997.

Seth Hettich, C.L. Blake, and Christopher J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html`.

David Hindle. Noun classification from predicate argument structure. In *Proceedings of the 28th annual meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburg, PA, 1990.

Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1) :177–196, 2001.

A. K. Jain, M. N. Murphy, and P.J Flynn. Data clustering : a review. *ACM Computing surveys*, 31(3) : 264–323, 1999.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, MA, 1997.

Frederik Jelinek, Lalit R. Bahl, and Robert L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21 :250–256, 1975.

Thorsten Joachims. Making large-scale svm learning practical. In Bernhardt Schölkopf, Christopher Burges, and Alex Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1999. The MIT Press.

Mark Johnson. Joint and conditional estimation of tagging and parsing models. In *Association for Computational Linguistics*, Toulouse, France, 2001.

Rosie Jones, Andrew McCallum, Kamal Nigam, and Ellen Riloff. Bootstrapping for text learning tasks. In *IJCAI-99 Workshop on Text Mining : Foundations, Techniques and Applications*, Stockholm, Sweden, 1999.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2000.

Adam Kilgariff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3) :333–347, 2003.

Dan Klein and Christopher D. Manning. Conditional structure versus conditional estimation in nlp models. Technical report, Dpt of Computer Science, Stanford, CA, 2002.

Hong-Kwang Jeff Kuo, Eric Fosler-Lussier, Hui Jiang, and Chin-Hui Lee. Discriminative training of language models for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 325–329, Orlando, FL, 2002.

Julian Kupiec, Jen Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73, 1995.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

Ray Lau, Ronald Rosenfeld, and Salim Roukos. Adaptative language modeling using the maximum entropy principle. In *Proceedings of the Human Language Technology Workshop*, 1993.

Ludovic Lebart, Alain Morineau, and Jean-Pierre Fénelon. *Traitement des données statistiques*. Dunod, Paris, 1979.

David D. Lewis. Data extraction as text categorization : an experiment with MUC 3 corpus. In *Proceedings of Third Message Understanding Conference (MUC 3)*, pages 245–255, San Diego, CA, 1991.

Lluís Màrquez. Machine learning and natural language processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, 2000.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

Andrew McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of the conference Uncertainty in Artificial Intelligence (UAI)*, Acapulco, Mexico, 2003.

Andrew McCallum and Kamal Nigam. A comparison of event models for Naive Bayes text classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.

Andrew McCallum, Dayne Freitag, and Fernando Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning (ICML)*, 2000.

Tony McEnery and Andrew Wilson. *Corpus Linguistics*. Edinburgh University Press, 1996.

Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1–8, Bergen, Norway, 1999.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

Ruslan Mitkov, editor. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003.

Kevin Murphy. *Dynamic Bayesian Networks : Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, 2002.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. In *Proceedings of NIPS*16*, pages Vancouver, Canada, 2002.

Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3) :103–134, 2000.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, 2002.

José Oncina, A. Castellanos, Enrique Vidal, and V. M. Jimenez. Corpus-based machine translation through subsequential transducers. In *Proceeding of the 3rd Conference on the Cognition Science of Natural Language Processing (CSNLP'94)*, Dublin, July 1994.

Patrick Pantel, Deepak Ravichandran, and Eduard Hovy. Towards terascale knowledge acquisition. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, Geneva, Switzerland, 2004.

Fernando Pereira, Naftali Tishby, and Lilian Lee. Distributional clustering of english words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics*, pages 183–193, Colombus, Ohio, 1993.

Jean-Marie Pierrel. *Ingénierie des Langues*. Hermès, Paris, 2000.

Vasin Punyakanok and Dan Roth. The use of classifiers in sequential inference. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems (NIPS*15)*, pages 995–1001. The MIT Press, 2001.

Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. PTR Prentice Hall, Englewoods Cliff, NJ, 1993.

Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey, 1995. Association for Computational Linguistics.

Deepak Ravichandran, Eduard Hovy, and Franz Josef Och. Statistical qa - classifier vs re-ranker : What's the difference ? In *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering–Machine Learning and Beyond*, Sapporo, Japan, 2003.

Reynar and Ardwait Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington, DC, 1997.

Jeffrey C. Reynar. *Topic Segmentation : Algorithms and Applications*. PhD thesis, University of Pennsylvania, 1998.

Ellen Riloff. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence, AAAI 99*, pages 474–479, Orlando, FL, 1999.

Brian Roark, Murat Saraclar, and Michael Collins. Corrective language modeling for large vocabulary asr with the perceptron algorithm. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 749–752, Montreal, Canada, 2004.

Mats Rooth. Two-dimensional clusters in grammatical relations. In *Proceedings of the symposium on representation and acquisition of lexical knowledge : polysemy, ambiguity and generativity*, Standford, CA, 1995.

Ronald Rosenfeld. Two decades of statistical language modeling : Where do we go from here ? *Proceedings of the IEEE*, 88(8), 2000.

Ronald Rosenfeld. *Adaptative Statistical Language Modeling : A Maximimum Entropy Approach*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1994.

Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu. Whole-sentence exponential language models : a vehicle for linguistic-statistical integration. *Computer, Speech and Language*, 15 :55–73, 2001.

Dan Roth. Learning in natural language : Theory and algorithmic approaches. In *Proceedings of the fourth Conference on Computational Language Learning*, pages 1–6, Lisbon, Portugal, 2000.

Dan Roth. Learning to resolve natural language ambiguities : a unified approach. In *Proceedings of the annual meeting of the American Association for Artificial Intelligence (AAAI)*, pages 806–813, Madison, WI, 1998.

Dan Roth. Learning in natural language. In *Proceeding of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 898–904, Stockholm, Sweden, 1999.

Antoine Rozenknop. *Modèles syntaxiques probabilistes non-gènératifs*. PhD thesis, Dpt. d'informatique, Ecole Polytechnique Fédérale de Lausanne, 2002.

Yves Schabes. Stochastic lexicalized tree-adjoining grammars. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 425–432, Nantes, France, 1992.

Ralf Schlüter, Wolfgang Macherey, Boris Müller, and Hermann Ney. Comparison of discriminative training criteria and optimization methods for speech recognition. *Speech Communication*, 34(3) :287–310, 2001.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2001. ISBN 0262194759.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1) : 1–47, 2002.

Terrence J. Sejnowski and Charles R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1 :145–168, 1987.

Satoshi Sekine. The domain dependence of parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 96–102, Washington, DC, 1997.

Kristie Seymore and Ronald Rosenfeld. Using story topics for language model adaptation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1987–1990, Rhodes, Greece, 1997.

Franck Smadja. Retrieving colloquations from texts : Xtract. *Computational Linguistics*, 19(1) :143–178, 1993.

Benjamin Taskar, C. Guestrin, and Daphne Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems (NIPS*19)*, Vancouver, Canada, 2003.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task : Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan, 2002. URL http://cnts.uia.ac.be/conll2002/proceedings.html.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning (ICML)*, 2004.

Peter D. Turney. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL-06)*, pages 313–320, Sidney, Australia, 2006.

Valtcho Valtchev, Jullian J. Odell, Phil Woodland, and Steve Young. MMIE training of large vocabulary recognition systems. *Speech Communication*, 44(4) :303–314, 1997.

Antal van den Bosch, Walter Daelemans, and Tom Weijters. Morphological analysis as classification : an inductive-learning approach. In Kemal Oflazer and Harald Somers, editors, *Proceedings of the Second International Conference on New Methods in Language Processing (NEMLAP II)*, pages 79–89, Ankara, Turkey, 1996.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

S. Vishwanathan and Alex Smola. Fast kernels for string and tree matching. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, Cambridge, MA, 2003. MIT Press.

Yiming Yang. An evaluation of statistical approaches to text categorization. Technical Report CMU-CS-97-127, Carnegie Mellon University, 1997.

David Yarowsky. Unsupervised word sense disambiguisation rivaling supervised methods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196, Cambridge, MA, 1995.

François Yvon. *Prononcer par analogie : motivations, formalisations et évaluations*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 1996.

John M. Zelle and Raymond J. Mooney. Learning semantic grammars with constructive inductive logic programming. In *Proceedings of the 11th National Conference of the American Association for Artificial Intelligence*, pages 817–822, Washington D.C., 1993.

Tong Zhang, Fred Damereau, and David Jonhson. Text chunking based on a generalization of winnow. *Journal of Machine Learning Research*, 2 :615–637, 2002.

Chapitre 2

Analogies

Les travaux présentés dans ce chapitre portent sur le développement[1] d'une famille d'apprentis, désignés sous le nom de systèmes *d'apprentissage par analogie*. Cet axe de recherche s'inscrit dans le prolongement direct de mon mémoire de thèse [Yvon, 1996], consacré aux méthodes de prononciation à base d'analogie. La motivation principale de ce travail est la recherche de mécanismes d'apprentissage qui soient spécifiquement adaptés aux spécificités des données linguistiques ; et partant plus performants que les apprentis statistiques « généraux » évoqués dans le chapitre 1, qui constituent aujourd'hui le courant dominant, y compris dans le domaine de la linguistique de corpus.

Mon hypothèse principale est que les données linguistiques présentent des caractéristiques tout à fait particulières, en particulier qu'elles s'organisent selon des *systèmes d'opposition* ; et que la prise en compte de ces caractéristiques via des biais d'apprentissage appropriés, comme ceux qu'exploite l'apprentissage par analogie, doit permettre de construire des apprentis performants *pour les tâches d'analyse linguistique*[2], voire de fournir des pistes de réflexion sur les mécanismes impliqués dans l'acquisition du langage.

En second lieu, cette recherche vise à développer des apprentis symboliques qui soient compatibles avec les descriptions linguistiques modernes et donc susceptibles d'incorporer tout ou partie des connaissances (linguistiques) accumulées sur ces données. Cette ambition distingue l'apprentissage par analogie tel que je l'envisage d'autres mécanismes d'apprentissage, dans lesquels l'intégration d'*a priori* linguistiques est plus difficilement envisageable.

Ces travaux de recherche, qui se sont déroulés pratiquement sans interruption depuis 1996, ont conduit au développement d'une famille originale d'apprentis, capables de traiter une large gamme de données linguistiques et obtenant des performances comparables, voire meilleures que celles des systèmes de l'état de l'art, en particulier sur des tâches d'analyse et de génération morphologique. Les développements formels les plus récents ont conduit à proposer des algorithmes efficaces implémentant les principes de l'apprentissage par analogie, utilisant le modèle des transducteurs finis.

J'ai choisi d'accorder dans ce mémoire une place relativement large à ces travaux, qui, pour certains n'ont jamais fait l'objet de présentation en français, ou n'ont été décrits que dans des rapports internes. L'organisation de ce chapitre respecte, pour l'essentiel leur chronologie : on y trouve en particulier les traces d'un va-et-vient entre expérimentation et modélisation. Les principes généraux qui sous-tendent les approches à

[1] Ces travaux ont été principalement poursuivis dans le cadre d'une collaboration, débutée il y a plus de 10 ans, avec V. Pirrelli, de l'Istituto di Linguistica Computazionale de Pise. Cette collaboration soutenue a donné lieu à de nombreux échanges et à des visites réciproques, qui ont considérablement influencé et enrichi ma réflexion sur l'analogie. Durant ces dernières années, ce travail s'est également nourri de discussions avec L. Miclet et A. Delhay de l'ENSSAT (Lannion). Les développements les plus récents doivent enfin beaucoup aux échanges avec N. Stroppa.

[2] Loin de moi, en particulier, l'idée que l'apprentissage par analogie pourrait résoudre de manière efficace toutes les tâches discutées dans le chapitre précédent.

base d'analogie en linguistique et en intelligence artificielle sont rappelés à la section 2.1. Ce rappel a une double visée : d'une part d'indiquer ce que cet apprenti emprunte à la tradition linguistique, et en particulier à la notion de paradigme. D'autre part, de permettre de me positionner vis-à-vis des recherches conduites en IA et en sciences cognitives autour du raisonnement par analogie. Cette longue section introductive s'achève par une présentation des motivations qui m'ont conduit à étudier ces modèles (section 2.1.4) ; cette présentation contient les éléments les plus originaux de ce début de chapitre.

La section suivante développe une première interprétation, déjà ancienne, de l'apprentissage par analogie (section 2.2). Je l'ai reprise ici, car dans une large mesure, elle reste toujours valide. Un aspect essentiel de cet apprenti y est souligné : sa mise en œuvre est presqu'intégralement spécifié par la définition de relations d'analogies entre les représentations linguistiques qu'il manipule (valeurs symboliques, séquences, structures de traits).

Je présente ensuite (section 2.3) quelques résultats d'expérimentations conduites sur deux tâches : la prononciation de mots inconnus (section 2.3.1) et l'analyse morphologique (section 2.3.2). Au terme de cette section, un certain nombre de limites de ce premier modèle sont discutées (section 2.3.3). Ces limites m'ont conduit en particulier à remettre en cause la définition de l'analogie sur les séquences.

La section 2.4 constitue donc un retour sur la formalisation de la notion d'analogie entre séquences. Une nouvelle définition est proposée, qui donne lieu à une implantation plus efficace des différents calculs nécessaires à la mise en œuvre de méthodes d'apprentissage par analogie, s'appuyant sur le modèle des transducteurs finis. Diverses extensions de ce modèle, actuellement en cours d'évaluation, sont enfin brièvement introduites.

Ce chapitre se termine par la présentation de quelques unes des perspectives ouvertes par ces travaux (section 2.5). Parmi les documents fournis en annexe de ce mémoire, les références [Stroppa and Yvon, 2005a,b] donnent un reflet fidèle des développements les plus récents de ce modèle.

2.1 De l'analogie

Le concept d'analogie possède une longue et riche histoire, dans des domaines aussi variés que la philosophie, les sciences cognitives expérimentales, l'intelligence artificielle et la linguistique. Sans viser à une présentation exhaustive de cette histoire, cette section revient sur ces diverses conceptions, partiellement contradictoires, de l'analogie, en examinant l'analogie en intelligence artificielle (section 2.1.1), puis en linguistique (section 2.1.2). Je présente ensuite diverses applications des modèles à base d'analogie en traitement automatique des langues (2.1.3), en m'arrêtant plus longuement sur quelques travaux significatifs.

2.1.1 L'analogie en intelligence artificielle et en sciences cognitives

Il est couramment admis que la capacité à effectuer des analogies est une compétence cognitive centrale chez l'humain (voir, pour un plaidoyer récent en faveur de cette thèse, [Hofstadter, 2001]). Cette compétence est supposée être utilisée dans un grand nombre de situations et pour résoudre des tâches variées de mémorisation/remémoration, de raisonnement (résolution de problème par identification d'un problème connu similaire, critique ou renforcement d'un argument sur la base d'exemples positifs ou négatifs), d'apprentissage (généralisation d'hypothèses) et de communication (compréhension de métaphores ou d'histoires, etc) [Haton et al., 1991, chap˙ 9]. Il n'est dès lors pas surprenant que de nombreux tests d'« intelligence » fassent la part belle à la résolution d'analogies, comprenant par exemple de multiples exercices de complétion de schémas analogiques [Hoffman, 1995]. Cette compétence analogique est volontiers associée aux capacités créatives de l'intelligence humaine, comme en témoigne la reconnaissance du rôle joué par les analogies entre domaines scientifiques dans le développement des sciences. La compréhension et la modélisation de

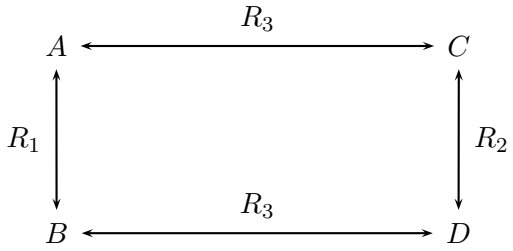

FIG. 2.1 – L'analogie comme un rapport de proportion

cette compétence sont donc des thèmes de recherche actifs dans les domaines de l'intelligence artificielle et des sciences cognitives (voir [Holoyak and Thagard, 1995, Gentner et al., 2001] pour des collections d'articles récents sur ces thèmes).

Cette section esquisse une discussion du concept d'*analogie* tel qu'il est traditionnellement utilisé dans le domaine de l'intelligence artificielle, en évoquant successivement l'analogie comme quatrième de proportion, l'analogie prédictive et les modèles de la perception analogique. Compte-tenu de la multiplicité et de la richesse des travaux conduits dans ce domaine, l'ambition de cette discussion est modeste : il s'agit, avant tout, d'exposer les idées essentielles de la démarche analogique, de distinguer les différents modes de l'analogie et de signaler les principales réalisations en la matière.

L'analogie proportionnelle

L'analogie *proportionnelle* constitue historiquement le premier mode de l'analogie [Lepage, 2003, p. 37-42]. Elle s'intéresse à des relations entre quatre termes A, B, C, D qui se glosent *A est à B ce que C est à D* (représenté graphiquement sur la figure 2.1). On utilisera la notation $A : B :: C : D$ pour dénoter ce type d'analogies. L'établissement d'un tel rapport induit des relations entre A et B d'une part, entre C et D d'autre part, qui ne valent toutefois que relativement à un contexte, qui est celui fourni par l'autre couple participant au rapport de proportionnalité.

Les propriétés classiques de l'analogie proportionnelle sont rappelées eg. dans [Lepage, 2001a] :

$$\forall A, B\ A : A :: B : B \tag{2.1}$$

$$\forall A, B, C, D\ A : B :: C : D \Rightarrow C : D :: A : B \tag{2.2}$$

$$\forall A, B, C, D\ A : B :: C : D \Rightarrow A : C :: B : D \tag{2.3}$$

La première de ces propriétés pose l'analogie comme généralisant l'égalité : il est toujours vrai qu'en positionnant deux objets identiques de part et d'autre du signe "est à" (:), on obtient une proportion licite. Les deux propriétés suivantes correspondent à des symétries de la relation : il est licite d'échanger les termes de part et d'autre du signe "comme" (::) ; il est également licite d'échanger l'ordre des termes impliqués dans la relation "est à".

La résolution *d'équations analogiques*, consistant à calculer le quatrième terme à partir des trois autres, est au programme de nombreux tests « d'intelligence », que ce soit dans le domaine des analogies verbales conceptuelles ou dans le domaine des formes géométriques. L'humain excelle à cet exercice et fait preuve d'une très grande flexibilité dans les processus d'identification de proportions analogiques. [Hoffman, 1995] donne un aperçu de cette flexibilité, en considérant la multiplicité des situations, correspondant chacune à des formes d'analogie différentes, que l'on peut obtenir en faisant varier le sens et la nature des relations R_1, R_2 et R_3 figurées sur le schéma 2.1.

La situation la plus simple est celle où R_1 est identique à R_2 (qu'il s'agisse d'une relation marquant une forme d'inclusion ou traduisant une dépendance fonctionnelle) et où R_3 traduit juste l'identité des relations.

L'analogie entre le soleil et la terre d'une part, et l'électron et le noyau atomique d'autre part, est un exemple de cette situation. La terre est en effet en rotation autour du soleil, de même que l'électron est en rotation autour du noyau : $R_1(= R_2)$ exprime cette double relation qui unit un satellite à son attracteur. L'analogie est établie par identification (via R_3) des forces qui provoquent ce mouvement de rotation. Dans cette configuration, le sens de l'analogie est indifférencié et elle fonctionne à l'identique dans les deux directions[3].

En s'appuyant sur de nombreux exemples dans le domaine des analogies verbales conceptuelles, Hoffman identifie une large palette de situations plus complexes, dans lesquelles, par exemple, R_1 et R_2 sont différentes, ou dans lesquelles R_3 n'est pas à proprement parler inversible, ou pour lesquelles l'analogie joue simultanément sur plusieurs dimensions sémantiques, ou encore pour lesquelles le niveau d'abstraction des relations entre objets est élevé.

La modélisation de cette capacité à résoudre des équations analogiques, si elle a attiré précocement l'attention des chercheurs (eg. [Evans, 1968]), s'est révélée relativement peu féconde : l'établissement de telles analogies ne relève, en première analyse, que des propriétés formelles des descriptions des objets entrant en relation et n'ouvre donc que peu de perspectives pour rendre compte de la créativité des processus analogiques [Indurkhya, 1989]. Nous reviendrons sur cette question des dynamiques analogiques ci-dessous, après avoir étudié un autre mode de l'analogie, les *analogies prédictives*.

Résolution de problème, analogies prédictives et raisonnement par analogie

[Carbonell et al., 1983] définit la résolution de problème par analogie de la manière suivante :

> Analogical problem solving consists of transfering knowledge from past problem solving episodes to new problems that share significant aspects with corresponding past experience and using the transferred knowledge to construct solutions to the new problems.
>
> *La résolution de problèmes par analogie consiste à transférer à de nouveaux problèmes des connaissances tirées de problèmes déjà résolus, pour autant que la nouvelle situation partage suffisamment de similarité avec des problèmes passés ; puis à utiliser ces connaissances pour construire des solutions aux nouveaux problèmes.*

Les analogies dont il est question ici n'impliquent donc que deux termes ; elles permettent de faire des inférences, justifiant leur appellation *d'analogies prédictives*. Si cette définition est suffisament large pour recouvrir une large gamme de situations, y compris des situations dans lesquels les problèmes source et cible ressortissent de domaines distincts (analogies inter-domaines), le domaine d'application par excellence du raisonnement par analogie est le raisonnement par cas [Kolodner, 1993, Aamodt and Plaza, 1994, Mitchell, 1997], dans lequel les problèmes mis en relation concernent un même domaine (domaine des recettes de cuisine, de la résolution de problèmes de bureautique, etc.).

La mise en œuvre de raisonnements par analogie ne requiert que la disponibilité d'une base de problèmes résolus, qui joue le rôle d'un modèle de la mémoire. Elle se décompose en plusieurs étapes principales :

– recherche dans la base d'exemples d'un problème S (dit problème *source*), partageant des caractéristiques voisines du problème T (le problème *cible*) en cours de résolution. Cette étape implique en fait deux mécanismes assez différents : d'une part, l'évocation de problèmes similaires à la cible, d'autre part la mise en correspondance des descriptions des sources candidates avec celle de la cible ;
– adaptation de la solution connue au nouveau problème, par un transfert de connaissances entre la source et la cible, qui conduit à *inférer* des propriétes de la cible. Durant cette étape, d'autres solutions peuvent également être mises à contribution, en conjonction avec la solution du problème source ;

[3] La symétrie n'est, naturellement, que formelle : dans les situations réelles de résolution de problèmes, l'analogie va toujours du connu vers l'inconnu. C'est encore le cas pour cet exemple : c'est historiquement la connaissance des lois qui régissent le mouvement des planètes qui a permis d'émettre des hypothèses sur la physique de l'atome, et pas l'inverse.

– additionnellement, la construction d'une analogie satisfaisante peut déboucher sur une étape d'induction, par construction d'un mécanisme de résolution plus direct s'appuyant sur des caractéristiques communes à la source et à la cible. Esquissons-en les grandes lignes : l'observation d'une ou de plusieurs occurrences de la conjonction des propriétés (P) et (Q) dans la base d'exemples conduit à induire une implication logique entre ces propriétés. Cette généralisation permet de supposer (Q) vraie chaque fois qu'une situation présente le caractère (P). On notera que dans ce mode d'*apprentissage par analogie*, l'analogie s'est, pour ainsi dire, perdue : il devient possible de déduire (Q) pour toute cible satisfaisant (P), et ce *indépendamment* de toute source.
– enfin, une résolution satisfaisante peut donner lieu à une étape de mémorisation du nouveau problème, d'autant plus nécessaire que cette mémorisation constitue un enrichissement réel de la base d'instances.

Les différentes étapes de cette démarche posent des problèmes difficiles et ont fait l'objet de multiples études (voir [French, 2002] et les références citées dans cet article). Du point de vue théorique, l'intérêt s'est toutefois largement focalisé autour de la question de l'identification des problèmes *pertinents* au regard de la nouvelle situation. Il apparaît en effet que le raisonnement par analogie se distingue de la déduction en ce que la validité de l'étape de transfert n'est pas logiquement fondée[4] conduisant à une interrogation sur la justification de telles inférences (voir eg. [Davis and Russel, 1987] pour une tentative pour proposer un cadre logique aux inférences analogiques).

Une justification possible consiste à présupposer qu'en général, des problèmes se ressemblant sous certains aspects ont également tendance à se ressembler sous d'autres aspects, incitant à rechercher les analogues parmi les problèmes les plus *similaires*. Il existe toutefois deux manières principales de concevoir la similarité entre problèmes, qui demande de distinguer une nouvelle acceptation du terme *analogie* : en plus du sens « technique » introduit ci-dessus, selon lequel l'analogie correspond à un rapport de proportion, ce terme possède en effet un sens plus « commun », dans lequel il est synonyme de *similitude*.

L'extraction de relations d'analogies-similitudes, c'est-à-dire *d'analogies de surface*, pose relativement peu de difficultés techniques et ne requiert que la définition d'une fonction de similarité entre problèmes, exemplairement fondée sur l'évaluation des caractéristiques que deux objets ont (ou n'ont pas) en commun. En ce sens, la méthode des k-plus-proches-voisins constitue une forme limite de raisonnement par analogie, applicable à des problèmes de catégorisation, dans lequel le raisonnement est réduit à son minimum.

Diverses observations ont toutefois conduit à suggérer que la résolution correcte de problèmes complexes passait par une mise en correspondance *structurelle* entre les problèmes source et cible. Comme l'exprime D. Gentner [1989, p.201] :

> (...) an analogy is a mapping of knowledge from one domain (the base) into another domain (the target), which conveys that a system of relations that holds among the base objects also holds among the target objects.
>
> (...) *une analogie est la mise en relation de connaissances entre un domaine (la source) et un autre domaine (la cible), qui exprime qu'un système de relations existant entre les objets du domaine source existe aussi entre les objets du domaine cible.*

Selon cette conception, les analogies les plus fécondes sont celles qui apparient des relations d'ordre supérieur dans les descriptions des problèmes cible et source (principe de *systématicité*), tout en respectant un principe de *cohérence structurelle* : si deux prédicats sont appariés, leurs arguments doivent également l'être. Alors que [Gentner, 1989] instancie ces principes de manière essentiellement syntaxique, en considérant des descriptions relationnelles des problèmes[5] (ramenant ainsi le problème de l'appariemment à un problème de matching de graphes), [Thagard et al., 1990] insiste sur l'importance de pondérer la qualité de l'appariemment par des critères *sémantiques* (ie. d'intégrer une proximité sémantique entre les descripteurs appariés) et *pragmatiques* (ie. de prendre en compte le but réellement poursuivi).

[4] Il est ainsi aisé de construire des exemples dans lesquels le raisonnement par analogie conduit à des conclusions erronées.
[5] Ou plutôt 'formel', pour ne pas entraîner de confusion avec l'autre sens, linguistique, de 'syntaxe'.

Le concept d'appariement structurel permet d'établir un lien entre analogie prédictive et la tradition classique d'analogie proportionnelle : l'analyse d'un appariemment entre deux descriptions fait ressortir des quadruplets de termes qui jouent deux à deux des rôles similaires dans la description des problèmes : ainsi la construction d'un appariemment entre les domaines « système solaire » et « atome » permet d'affirmer par exemple que « la terre est au soleil comme l'électron est au noyau atomique ». Par ce biais, l'analogie prédictive retrouve la tradition classique de l'analogie comme rapport de proportionnalité.

Il semble donc qu'il existe deux visions distinctes de l'analogie prédictive : l'une mettant en jeu des calculs de similarité, l'autre des appariemments structurels et/ou fonctionnels plus abstraits. Il serait pourtant réducteur d'appréhender cette dichotomie sous la forme d'une opposition entre une version « simpliste » et une version plus « réaliste » des phénomènes d'association analogiques ; le rapport entre ces deux mécanismes doit plutôt s'envisager sur le mode de la complémentarité, tant sur le plan cognitif que sur le plan de la modélisation informatique [Forbus et al., 1994]. Il semble, en effet, que la capacité d'un sujet humain à extraire des analogies de plus en plus complexes (ie. impliquant des appariemments structurels) dans un domaine augmente avec la connaissance de ce domaine. Comme le note également [Itkonen and Haukioja, 1997], la production d'analogies se manifeste comme un continuum entre, d'un côté, des généralisations spontanées et inconscientes, telles que celles qui se forment par exemple lors de l'acquisition du langage, et de l'autre, des raisonnements conscients utilisés pour résoudre des problèmes nouveaux, ou encore dans le cadre de la découverte scientifique. Il semble enfin que, du point de vue algorithmique, le passage à l'échelle demande de faire conjointement appel aux deux types de mécanismes liés à ces deux conceptions : le calcul d'analogies de surface, moins coûteux algorithmiquement, servant alors à filtrer les objets susceptibles de participer à une analogie structurelle.

La conception de systèmes cognitivement plausibles implémentant des analogies prédictives a donné lieu à de multiples tentatives de formalisation et de mise en œuvre, chacune s'attachant à modéliser plus particulièrement certains aspects du calcul des appariemments analogiques : la préférence pour les analogies structurelles, pour celles qui impliquent également une proximité sémantique ou pragmatique, pour celles qui exploitent une connaissance du domaine, pour celles qui intègrent des contraintes liées à la limitation de la mémoire de travail... Citons, en particulier, [Falkenhainer et al., 1990, Thagard et al., 1990, Keane et al., 1994, Kokinov, 2001] pour diverses réalisations emblématiques des années 90 : SME, ACME, IAM et AMBR, ainsi que [French, 2002] pour un état de l'art récent. Parallèlement à ces études, qui restent marquées par la tradition symboliste en IA, des travaux plus récents proposent des modèles d'appariemment structurels utilisant des représentations sub-symboliques distribuées (eg. [Plate, 2000, Holoyak and Hummel, 2001]).

Analogie et perception

Si les modèles du raisonnement par analogie introduits à la section précédente ont conduit à de nombreuses avancées théoriques et pratiques, ils ont également fait l'objet de vives critiques, en particulier [Indurkhya, 1989, Chalmers et al., 1992]. Selon ces auteurs, la principale faille de ces modèles provient de ce que (i) ils considèrent des relations entre des représentations statiques, supposées préexister à l'établissement de l'appariemment entre source et cible ; (ii) ils privilégient des appariements formels (syntaxiques) entre ces représentations. Ce faisant, ils passent à côté de la véritable source de créativité des phénomènes analogiques, qui font émerger et incitent à interpréter des ressemblances entre cible et source *qui n'étaient pas originellement présentes*. Dit autrement, la mise en relation d'une source et d'une cible conduit à jeter un œil neuf sur la cible, qui se trouve réanalysée dans des termes fournis par la source. La modélisation de l'analogie doit donc d'emblée intégrer la manière dont la co-perception de deux objets *modifie dynamiquement leurs représentations*, conduisant à faire émerger des représentations unifiées des situations. Pour [Indurkhya, 1989], les métaphores[6] ressortissent de ce type d'analogie : un emploi métaphorique force la réinterprétation d'un terme dans un contexte (environnement) étranger (celui de la métaphore), le dotant

[6] Nuançons, à la suite de [Lepage, 2003, p. 40] : certaines métaphores.

ainsi d'éléments de significations nouveaux.

Toujours selon Indurkhya, cette faiblesse empêche les modélisations traditionnelles du raisonnement par analogie de percevoir des analogies entre des situations dont les représentations sont très éloignées, alors même qu'elles sont les plus instructives pour améliorer la compréhension de la source. Cette observation le conduit à distinguer plus précisément deux types d'analogies : sont strictement syntaxiques[7] les proportions qui n'impliquent pas de réinterprétation des objets considérés :

> Thus, we associate a description with every object, and call it the *natural description* (...) Now any proportional analogy relation that can be comprehended from the natural description of the its objects is syntactic. On the other hand, if a redescription of at least one object is necessary, then the analogy relation is interpretative.
>
> *En conséquence, nous associons une description avec chaque objet, et l'appelons description naturelle (...) Toute analogie proportionnelle qui peut être appréhendée depuis la description naturelle des objets est syntaxique. D'un autre côté, si une redescription est nécessaire, alors la relation d'analogie est interprétative.*

À l'inverse, si l'établissement de relations de proportion inclut une forme de réinterprétation dynamique des termes en relation, alors on a affaire à une analogie (proportionnelle) *interprétative*.

L'importance des processus de réanalyse des objets entrant dans une analogie est également attestée pour les analogies proportionnelles : [Hofstadter and Mitchell, 1995] s'intéresse précisément à ces situations, donnant lieu au système Copycat, qui résout des analogies proportionnelles en construisant dynamiquement un cadre interprétatif commun pour les objets participant à la relation analogique. [Dastani et al., 1997, Schmid et al., 2003] proposent d'autres modèles intégrant des capacités de réanalyse pour la résolution d'analogies proportionnelles entre chaînes de caractères. Ces deux modèles présentent un cadre algébrique, fondé sur une famille d'opérateurs primitifs (l'itération, la symétrie, l'égalité, la précédence, etc.) qui permettent de doter chaque chaîne d'une multitude d'interprétations, certaines plus « naturelles » que d'autres. L'établissement d'un rapport de proportion conduit à réviser l'interprétation des chaînes analogues, en favorisant des analyses qui conduisent à une interprétation unificatrice des quatre chaînes en présence.

En résumé : quelques lignes de démarcation

Cette étude rapide de l'histoire de l'analogie en Intelligence Artificielle a conduit à rappeler que les processus analogiques jouent un rôle central dans la cognition humaine. Ils permettent d'établir des similarités entre deux situations, fondées sur la mise en relation (appariemment structurel) de termes ou de relations entre termes intervenant dans la représentation de ces situations. Cette mise en relation se traduit par l'identification de rapports de proportions, dont le trait commun est d'impliquer au moins quatre éléments.

Au-delà de ce socle commun, des conceptions multiples de l'analogie ont été proposées et étudiées, donnant lieu à une série d'oppositions : analogies proportionnelles vs. analogies prédictives, analogies de surface vs. analogies structurelles ; analogies syntaxiques vs. analogies interprétatives, analogies inter-domaine vs. analogies intra-domaine, analogies spontanées vs. analogies conscientes, etc. Ces oppositions permettront de mieux positionner mes propres travaux en la matière.

2.1.2 L'analogie en linguistique

Le concept d'analogie possède également une longue histoire dans la tradition linguistique. Quelques-uns des principaux moments de cette histoire sont rappelés dans cette section, permettant au passage de souligner

[7] Ce cas inclut naturellement des analogies qui pourraient, sous un autre aspect, être qualifiées de sémantiques : par exemple celle que l'on pourrait établir entre les concepts *jument* et *poulain*, d'une part, et *poule* et *poussin*, d'autre part, sur la base d'une sémantique à base de traits impliquant les distinctions entre enfant et adulte, etc.

à nouveau l'étendue des significations de ce terme. [Lepage, 2003], ainsi que [Itkonen and Haukioja, 1997, Lavie, 2003, Dal, 2003] constituent des sources d'information essentielles pour se faire une vision plus complète de cette histoire.

Dès l'antiquité, l'analogie (proportionnelle) est proposée comme un principe de description du savoir linguistique des locuteurs, particulièrement adapté à la description des systèmes flexionnels. L'examen de la distribution des flexions nominales ou verbales du grec ou du latin fait, en effet, apparaître des séries congruentes, dont l'organisation peut être représentée au moyen de tables croisant lemmes et propriétés morpho-syntaxiques. Pris deux à deux, les termes de ces séries se présentent sous la forme de rapports de proportion. Ainsi, en latin :

(3) *amabam* : *amabat* :: *legabam* : *legabat*
(*j'aimais* : *il aimait* :: *je lisais* : *il lisait*)

Dans cette conception, encore vivace à l'époque classique, l'analogie est un principe *statique* de description du *régulier* (morphologique) et s'oppose à *l'anomalie*. Ainsi, le grammairien Arnauld, cité par [Lavie, 2003], exprime clairement l'idée que l'analogie est la règle :

> C'est une maxime que ceux qui travaillent sur une langue vivante doivent toujours avoir devant les yeux, que les façons de parler qui sont autorisées par un usage général et non contesté, doivent passer pour bonnes, encore qu'elles soient contraires aux règles et à l'analogie de la langue ; mais on ne doit pas les alléguer pour faire douter des règles et troubler l'analogie, ni pour autoriser, par conséquent, d'autres façons de parler que l'usage n'aurait pas autorisé.

Comme le note G. Dal, [2003], l'analogie va demeurer pendant plusieurs siècles un principe permettant de justifier la construction de nouveaux mots sur la base de mots existants, ce qu'elle illustre par ces extraits, tirés respectivement de la première édition et de la sixième édition du Dictionnaire de l'Académie :

> « le mot 'ambitionné' est formé par analogie d'ambition, comme 'passioné' est formé de 'passion'. »
>
> « [l]es mots nouveaux ne peuvent guère s'introduire qu'à l'aide de l'analogie ».

Un second moment important de cette histoire se produit au tournant du 20e siècle, avec la reconnaissance de l'analogie comme un principe *dynamique*, qui permet d'expliquer l'apparition de nouvelles formes. Ce renouveau de l'analogie s'opère alors que l'intérêt des linguistes est tourné vers la linguistique diachronique et l'identification de familiarités entre langues ; ce travail s'opérant par la mise à jour de *lois phonétiques*, qui permettent de retracer les évolutions historiques, à partir d'une racine commune, de lexèmes appartenant à des langues différentes. Saussure [1916], en particulier, consacre un chapitre entier à l'analogie, pour montrer comment la dynamique des réfections analogiques permet de produire des formes plus *régulières* et vient ainsi contrebalancer l'effet des changements phonologiques. Il analyse ainsi, dans un exemple resté fameux, l'évolution de la paire latine (*honōs*, *honosem*), évoluant par changement phonétique en (*honōs*, *honorem*), puis en (*honor*, *honorem*). Ce second changement est analysé comme résultant de la pression d'alternances régulières telles que (*orator*, *oratorem*), alternance qu'on observe plus généralement dans la déclinaison des agentifs masculins. Le façonnage d'*honor* apparaît ainsi comme le résultat de la résolution d'une équation analogique : *orator* : *oratorem* :: *honor* : *?*.

Ces analogies sont effectivement créatives (interprétatives), dans la mesure où l'influence de couples d'analogues pousse à la *réanalyse* de formes connues : ainsi l'existence du verbe *somnoler* pousse-t-il à réanalyser *somnolent* comme *somnol+ent*, alors que l'étymologie correcte suggère un découpage *somn+olent* [de Saussure, 1916, p. 234]. Plus près de nous, pensons à des formes telles que *workholic*, induisant une réanalyse de *holic* comme un suffixe de *alcoholic*.

Retenons de ce renouveau moderne de l'analogie, la confirmation par Saussure et plusieurs de ses contemporains, de la grammaticalité et de la productivité des procédures analogiques.

[Bloomfield, 1933, p. 275 et suivantes] étend le principe descriptif de l'analogie (proportionnelle) à la description syntaxique :

> « Les analogies régulières d'un langage sont des habitudes de substitution ».

C'est toutefois plus la substituabilité que l'analogie qui intéresse Bloomfield, qui tire de la possibilité pour des formes ou des groupes d'alterner en certaines positions, un principe permettant à la fois d'identifier des unités linguistiques (morphèmes et syntagmes) et de généraliser leurs arrangements ; la description de ces unités et de leur combinatoire va, dès lors, occuper une attention croissante de la part des linguistes.

On sait que, dans l'accomplissement de ce programme, la linguistique générativiste moderne a largement discrédité le concept d'analogie : une citation parmi d'autres, tirée de [Chomsky, 1975] :

> « (...) des notions comme l'analogie ne le [le scientifique libre de toute idéologie] mènent pas loin dans l'étude des capacités humaines, au moins dans le domaine du langage ».

Un des exemples donnés par Chomsky [Chomsky, 1975, p. 140] pour discréditer l'analogie (vue comme un principe d'induction) est la comparaison entre (4a) et (4a), qui, quoique très similaires, s'analysent et se comprennent de manière fort différente :

(4) a. *John's friends appeared to their wives to hate one another*

b. *John's friends appealed to their wives to hate one another*

Comme le note [Itkonen and Haukioja, 1997], cet argument n'est pas recevable, en particulier parce que l'analogie dont parle Chomsky n'est pas l'analogie au sens d'un rapport de proportion (similitude structurelle), mais une similarité « physique » entre des formes de surface[8]modèles plus conformes au concept d'analogie comme proportion (ou comme appariement structurel), il devient tout à fait possible de traiter correctement les énoncés jugés hors de la portée de l'analogie par Chomsky [Itkonen and Haukioja, 1997, Lavie, 2003]. Cette réhabilitation par Itkonen de l'analogie est illlustrée par un modèle symbolique simple capable de construire et d'analyser des énoncés complexes tels que (4b) et (4b). Elle s'accompagne d'une tentative plus large de justifier de l'utilisation de tels modèles en linguistique et pour corrollairement réfuter les postulats principaux de la tradition chomskienne en linguistique (désintérêt pour les processus de découverte, accent mis sur les modèles de compétence et en particulier sur la syntaxe, dont l'autonomie postulée est une des pierres angulaires du système chomskien, etc.). Selon Itkonen, les modèles analogiques permettent au rebours de faire le lien entre la capacité de langage et d'autres capacités mentales, de fournir un modèle du développement de cette capacité, de servir de modèles aux *processus* de construction et de compréhension d'énoncés nouveaux et finalement d'ouvrir les voies vers une réunification de la linguistique.

Cette réhabilitation de l'analogie en linguistique prend un tour évident dans le domaine de la morphologie. En ce qui concerne la morphologie flexionnelle, de nombreux auteurs, à la suite de Matthews [1974], s'accordent pour reconnaître un rôle central au concept de *paradigme*, et envisagent la production de formes fléchies d'un lexème sous l'angle de l'association d'une forme 'principale' du lexème avec un paradigme exemplaire, association qui permet de reconstruire toutes les autres formes *par analogie*. Comme l'expose [Blevins, 2006] :

> « Matching the principal part of an item with against cells of an exemplary paradigm establishes a correspondence between principal parts and their counterparts in the exemplary paradigm. New forms of an item are deduced by extending this correspondence analogically to other cells. »

En ce qui concerne la morphologie dérivationnelle, l'hypothèse de la 'Morphologie Paradigmatique' tend également à rassembler un nombre croissant de défenseurs. L'extrait suivant, tiré de [Booij, (to appear], en donne l'idée principale :

[8]Dans le passage pré-cité, Chomsky prend soin, dans un premier temps, de caractériser les analogies qu'il récuse et qui sont les analogies *apparentes* (« ainsi, bien que (4a) et (4b) se ressemblent beaucoup, les locuteurs ne les comprennent pas moins très différemment, sans tenir compte des ananalogies *apparentes* »). Il abandonne toutefois dans la suite toute distinction (« Ce sont là des exemples typiques qui conduiraient au rejet d'une idée traditionnelle : l'analyse de la langue pourrait se fonder sur les notions d'analogie et de généralisation. »).

> Word-based morphology is the idea that morphology is primarily a set of systematic correspondences between the forms and meanings of the words of a language. These systematic correspondences can also be applied to new cases, and then we speak of productive morphological processes that create words from words (Schultink 1961). That is, the source of morphology is the network of paradigmatic relations beween the existing words of a language. (...) This [the paradigmatic view of morphology] position implies that it is the word that forms the basis of morphological operations, and that morphology cannot be defined as the concatenation of morphemes into words.

Un intérêt évident de cette hypothèse est qu'elle permet de donner une vue unifiée des procédés de formation de de mots en regroupant sous la banière de l'analogie des mécanismes allant du plus régulier (affixation) au plus marginal (mot-valises, fausses étymologies, etc). Plus profondément, comme l'analyse [Dal, 2003, Booij, (to appear], elle permet de proposer un modèle explicatif qui résoud certaines des apories des modèles syntagmatiques en morphologie, comme les substitutions de suffixes ou de constituant, les formations régressives (*back-formations*), les paradoxes de parenthésage et les cas de syncrétisme.

2.1.3 L'analogie en traitement automatique des langues

Cette section dresse un rapide panorama de l'utilisation de techniques dites « à base d'analogies » pour des tâches classiques du traitement automatique des langues, permettant de constater que l'ambivalence de terme en IA et en linguistique se retrouve largement en traitement automatique des langues. Cette section s'achève par une présentation un peu plus longue des travaux les plus aboutis sur l'analogie comme rapport de proportion, à savoir ceux conduits par l'équipe d'Yves Lepage.

Une multiplicité de modèles

L'analogie en traitement automatique des langues est le plus souvent entendue dans le sens général de *similitude de surface*, quelques travaux relativement marginaux faisant plus explicitement référence à la recherche et à l'utilisation de rapports de proportionnalité.

Au fondement de ces travaux très variés, trois visées très différentes, qui permettent d'en dresser une première typologie :

– d'un côté, des travaux inspirés par un certain nombre de résultats expérimentaux en psychologie cognitive, qui tendent à mettre en évidence, pour certaines tâches spécifiques (lecture, accès lexical, morphologie, etc), le caractère analogique (c'est-à-dire non gouverné par des règles) des opérations réalisées par le cerveau. L'objectif est de proposer des modèles pour ces processus et d'alimenter ainsi, dans la querelle qui anime les sciences cognitives entre partisans et adversaires des règles, le camp des défenseurs de l'empiricisme. On trouvera, dans [Dedina and Nusbaum, 1991, Damper and Marchand, 1998] pour la tâche de prononciation et dans [Ling and Marinov, 1993, Marcus, 1995, Plunkett and Marchman, 1996] pour la morphologie, un échantillon représentatif de cette démarche.
– une seconde famille de travaux vise, en se fondant non plus sur des résultats de psychologie cognitive, mais sur les arguments évoqués à la fin de la section 2.1.2, à proposer des mécanismes de calcul pour modéliser les formes d'apprentissage, de connaissance et de raisonnement impliqués par l'activité langagière ; permettant ainsi de se dispenser de l'hypothèse de prédispositions s'exprimant par des systèmes de règles pour modéliser le savoir linguistique. Les travaux de R. Skousen [1989, 2002], de D. Jones [Jones, 1996], de Y. Lepage [2003] ou encore de R. Lavie [2003] se situent dans cette ligne de recherche, avec, dans le cas de Skousen, l'ambition de proposer un modèle qui soit également compatible avec des données de psychologie expérimentale ;
– enfin, un certain nombre de travaux s'inscrivent dans une démarche plus pragmatiques et s'attachent à montrer que, nonobstant les arguments issus des deux traditions précédentes, les approches à base d'analo-

gie doivent être encouragées pour les performances qu'elles permettent d'atteindre pour une vaste gamme d'applications linguistiques. C'est le cas en particulier des travaux de W. Dealemans et de son équipe (voir e.g. [Daelemans et al., 1996, van den Bosch, 1997, Zavrel et al., 1997, van den Bosch and Daelemans, 1999, Veenstra et al., 2000, Zavrel and Daelemans, 2003]) ; mais également de ceux de [Coker et al., 1990, Bagshaw, 1998, Damper et al., 1998] ; ainsi que des quelques applications en traitement automatique des langues du raisonnement par cas (eg. [Golding, 1991, Cardie, 1997]). C'est principalement des considérations similaires qui m'ont conduit à considérer avec attention des techniques d'apprentissage par analogie.

L'analogie comme similarité Si ces travaux poursuivent des buts variés, la majorité d'entre eux ne partagent pas moins une prémisse essentielle : l'analogie y est le plus souvent envisagée comme une similarité de forme (analogie de surface). Citons, par exemple, [Jones, 1996, p. 4] :

> « (...) analogy here refers to the general process of deriving information about some new piece of language by comparison with known pieces of language (...) »
>
> »(...) *le mot 'analogie' fait ici référence au procedé général de dérivation d'une information concernant un échantillon linguistique nouveau par comparaison avec des échantillons connus (...)* »

Ce point de vue très majoritaire permet d'expliquer les raisons pour lesquelles des techniques d'apprentissage très diverses[9] se retrouvent enrôlées sous la banière de l'analogie : variantes des k-plus proches voisins, apprentissage de règles, arbres de décision, réseaux neuronaux, etc., qui n'ont de commun que de mettre en œuvre l'idée d'une généralisation à partir de données.

De ces travaux, deux principes essentiels, explicités eg. dans [Daelemans et al., 1999] méritent d'être retenus :

– la connaissance linguistique se construit en extension, par accumulation progressive d'exemples, qui *tous*, y compris les plus exceptionnels, contribuent à la capacité de généralisation du système ;
– la généralisation s'opère toujours dans le contexte particulier d'une nouvelle instance à traiter ; elle ne mobilise qu'une partie restreinte des connaissances disponibles, exemplairement les exemples les plus proches de la nouvelle instance.

L'analogie comme proportion Par comparaison, relativement rares sont les travaux qui envisagent l'analogie comme un rapport de proportionnalité : [Blin and Edgington, 2000, Blin and Miclet, 2000] envisage l'apprentissage par analogie de structures prosodiques, demandant la recherche en corpus de rapports de proportions entre arbres étiquetés, les principaux résultats obtenus concernent toutefois plus les aspects algorithmiques du calcul des similarités entre arbres ; [Hathout, 2001] recherche des rapports de proportions analogiques entre lemmes parmi les membres d'une même famille dérivationnelle pour inférer automatiquement des shémas constructionnels qui sont utilisés pour enrichir un graphe de relations morphologiques. Les plus aboutis de ces travaux sont toutefois dûs à Y. Lepage ; la section qui suit leur est intégralement consacrée.

Cette section serait incomplète sans la mention des travaux récents de P. Turney [2003, 2005, 2006], qui se distingue clairement, dans ses moyens, des approches précédentes. Ce travail s'intéresse certes à la résolution d'analogies proportionnelles, en particulier d'analogies verbales conceptuelles telles que celles qui étaient utilisées dans les tests SAT[10].Turney, à la différence des auteurs précédents, utilise toutefois pour ce

[9] Cette confusion rend d'ailleurs problématique l'émergence d'une conception cohérente de l'analogie - ainsi par exemple, les travaux déjà mentionnés de R. Skousen, qui figurent parmi les plus influents sur ce thème, n'ont que très peu à voir avec la vision de l'analogie comme rapport de proportion, ni avec la vision de l'analogie comme mise en relation de structures qui prévalent, notamment en IA.

[10] Le *SAT Reasoning Test*, autrefois appelé *Scholastic Aptitude Test*, est un des tests utilisé pour évaluer le niveau des étudiants

faire non pas des méthodes à base d'analogie, mais des techniques (modèle vectoriel, LSA) inspirées de la recherche d'information. J'aurai l'occasion, dans le prochain chapitre, de resituer ce travail en discutant une contribution voisine consacrée à l'acquisition automatique de paraphrases (voir la section 3.1).

Les algorithmes d'Y. Lepage

Le travail d'Y. Lepage vise à modéliser et à exploiter dans un cadre de traitement automatique des langues les analogies proportionnelles entre séquences de symboles. [Lepage, 1998] introduit une *définition calculatoire* de l'analogie sur des séquences, en présentant un algorithme de calcul d'analogie sur des mots formels, fondé sur l'utilisation d'une mesure de dissimilarité déduite de la distance d'édition classique. Pour résoudre une équation telle que : $x : y :: z :?$, cet algorithme procède en deux étapes :

– calcul d'un alignement optimal, au sens de cette dissimilarité, d'une part entre x et y, d'autre part entre x et z. Cette étape reprend les algorithmes classiques de calcul d'une distance d'édition entre séquences par programmation dynamique [Wagner and Fischer, 1974] ;
– réconciliation *heuristique* de ces deux alignements, guidée par la contrainte que chaque symbole de x s'aligne avec exactement un symbole émanant soit de y, soit de z ;

Diverses applications de cet algorithme sont présentées dans [Lepage, 1999b], auquel le travail sur la morphologie flexionnelle présenté à la section 2.3.2 ; ainsi que dans [Lepage, 1999a], qui s'intéresse aux analogies entre arbres, préalablement linéarisés par parenthétisation. Ce travail algorithmique a fortement influencé mes réflexions méthodologiques, comme il apparaîtra clairement à la section 2.4, dans laquelle cet algorithme sera (i) ré-exprimé à l'aide de transducteurs finis, permettant de mieux apprécier son caractère non-déterministe ; (ii) généralisé de diverses manières, permettant de définir des notions moins contraintes et possiblement graduelles de l'analogie sur les mots.

[Lepage, 2000, 2001b] propose une seconde définition, cette fois-ci *axiomatique*, de l'analogie sur les séquences, à partir de laquelle il dérive un mécanisme *génératif* pour ce qu'il appelle les *langages analogiques*. Partant des axiomes généraux de l'analogie (symétrie, égalité des moyennes...), l'idée principale consiste à spécifier le comportement de l'opération de concaténation par rapport à la propriété d'analogie selon :

$$x_1 : y_1 :: z_1 : t_1 \text{ et } x_2 : y_2 :: z_2 : t_2 \Rightarrow \begin{cases} x_1x_2 : y_1y_2 :: z_1z_2 : t_1t_2 \\ x_1x_2 : y_1y_2 :: z_2z_1 : t_2t_1 \\ x_1x_2 : y_2y_1 :: z_1z_2 : t_2t_1 \end{cases} \tag{2.4}$$

Étant donné un *ensemble générateur* M de couples de mots, une relation, notée $\vdash_M$, de dérivation immédiate entre mots est alors définie par : $x \vdash_M y$ ssi $\exists (z,t) \in M$ telle que $x : y :: z : t$. Si $\vdash^\star_M$ est la clôture transitive de cette relation, le langage analogique dérivé par M du langage L_0 est alors :

$$\{y, \exists x \in L_0, x \vdash^\star_M y\}.$$

S'appuyant sur cette définition, [Lepage, 2001a] démontre que certains langages hors-contexte ou faiblement sensibles au contexte sont des langages analogiques. Si je ne le suis pas dans ces derniers développements (voir [Yvon, 2003] pour une discussion), retenons en revanche de ce dernier travail la notion de langage analogique, qui sera reprise à la section 2.4.

2.1.4 Pour une approche analogique

Au terme de ce tour d'horizon, cette section présente des arguments supplémentaires qui justifient à mes yeux l'utilisation des méthodes à base d'analogie en traitement automatique des langues. À l'origine de mon

à leur entrée dans les universités américaines. Jusqu'en 2005, il incluait une section consacrée à la résolution d'analogies verbales. Voir http://en.wikipedia.org/wiki/SAT#The_SAT_and_IQ.

intérêt pour ces méthodes, se trouve une analyse critique de l'utilisation de méthodes classificatoires pour des tâches d'analyse linguistique[11], développée dans mon travail de thèse [Yvon, 1996, p. 74 et suivantes] et poursuivie dans [Pirrelli and Yvon, 1999b]. Après avoir rappelé ces critiques, je présente les raisons pour lesquelles il me semble possible de leur répondre en prennant appui sur les notions de *commutation* et de *paradigme*. Au terme de cette discussion, j'essaierai de positionner ma démarche vis à vis d'autres approches analogiques en traitement automatique des langues, en particulier celles évoquées à la section précédente.

Retour sur l'activité classificatoire

D'un point de vue linguistique[12], la mise en œuvre d'une approche classificatoire en traitement automatique des langues, décrite en particulier dans la section 1.2.2, repose crucialement sur deux hypothèses discutables : (a) l'hypothèse *atomistique*, qui postule l'existence d'unités qui constituent les atomes pour un niveau donné de description (les lettres, les morphèmes, les mots, les syntagmes) et qui font l'objet du processus de classification et (b) l'hypothèse de *compositionalité* : qui envisage l'interprétation d'une séquence d'atomes comme découlant de la composition des interprétations des atomes individuels.

Dans le contexte de tâches d'analyse linguistique, les hypothèses (a) et (b) ne constituent pourtant qu'une approximation dont l'appréciation des limites permet de questionner la mise en œuvre de méthodes classificatoires. En particulier, ces hypothèses semblent fragilisées par les trois observations suivantes.

En premier lieu, l'ambiguïté de la délimitation des atomes, dont les frontières sont, le plus souvent, non explicites[13]et doivent être reconstruites (voir la discussion sur la segmentation à la section A.0.6). Il est également important de noter que ces frontières sont instables entre niveaux, comme l'illustrent les exemples suivants :

– non concordance des frontières entre l'organisation prosodique et l'organisation orthographique d'un énoncé : une syllable peut impliquer (par exemple en cas de liaison) des phonèmes "issus" de plusieurs mots ; un mot phonologique peut s'étendre sur plusieurs mots orthographiques ; la structure prosodique est enfin non congruente à la structure syntaxique (voir, par exemple, pour le français les travaux de [Delais, 1995, Lacheret-Dujour and Beaugendre, 1999]).
– non concordance entre l'organisation prosodique ou l'organisation orthographique et l'interprétation sémantique d'un dérivé ou d'un composé, en particulier les paradoxes de parenthésage [Spencer, 1991] ;
– non concordance entre l'organisation syntaxique d'un énoncé et l'interprétation qui peut en être donnée. L'existence de mots composés, d'idiomes, etc. en est la trace la plus évidente ; un autre exemple est fourni par les problèmes de portée des quantifieurs, qui sont à la fois internes à un groupe nominal tout en ayant une portée sur l'ensemble d'une proposition.

En second lieu, l'ambiguïté dans la détermination de l'interprétation d'un atome. À tous les niveaux de représentation (orthographique, morphologique, syntaxique, sémantique), les cas de polysémie sont légion. Ce point est largement documenté dans la littérature et est au fondement même des approches probabilistes en linguistique ; il ne semble donc pas devoir être développé plus avant ici.

En dernier lieu, l'hypothèse de compositionnalité, enfin, semble une hypothèse trop forte pour rendre compte du fonctionnement des entités linguistiques. De nouveau, l'on dispose de multiples exemples de violation de l'hypothèse de compositionnalité : dans la correspondance orthographique/phonétique, en analyse morphologique, dans l'interprétation sémantique de syntagmes... Notons, de surcroît, qu'il existe une graduation

[11]En particulier les tâches de transcodage, dans lesquelles les entrées et sorties sont des séquences de symboles. Une ligne de critique plus radicale des modèles classificatoires consiste à remettre en cause la possibilité de construire des catégories opératoires pour la description linguistique (tout particulièrement en syntaxe et en sémantique, cf. à nouveau [Lavie, 2003, p. 275 et suivantes]) : ma critique se développe toutefois sur un plan différent.

[12]La discussion de cette section et de la suivante suit, pour l'essentiel, [Pirrelli and Yvon, 1999a].

[13]Pour autant même qu'elles puissent être identifiées : on sait la difficulté que pose aux approches classiques en morphologie l'existence de morphèmes « porte-manteau » véhiculant une multiplicité d'éléments d'interprétation.

dans la manière dont la stricte compositionnalité est rompue, qui fait que la question ne peut être ramenée simplement à une question de segmentation. Ces phénomènes sont loin d'être anecdotiques et reflètent, dans une large mesure, une tension entre, d'une part, un fonctionnement entre niveaux de description qui reste globalement compositionnel et d'autre part une relative autonomie des *systèmes* que forment, pour un niveau de représentation donnée, les entités linguistiques, qui les conduisent à évoluer diachroniquement vers une organisation plus régulière.

La démarche classificatoire n'apporte véritablement de réponse qu'à la seconde de ces objections (levée d'ambiguïtés), puisqu'elle conduit à calculer *en contexte* l'association entre unités et interprétations ; le contexte (au sens large) apparaissant alors comme le moyen de lever les ambiguïtés d'interprétation. C'est le fondement de l'utilisation pour des tâches d'analyse linguistique de méthodes d'apprentissage classiques, en particulier les méthodes statistiques longuement évoquées dans le chapitre précédent. Ces objections semblent suffisament sérieuses pour justifier l'utilisation de méthodes alternatives. Le paragraphe qui suit s'attache à montrer que les méthodes à base d'analogie permettent de répondre à quelques unes de ces objections.

Paradigmes, commutation, analogie

L'approche préconisée ici vise à s'affranchir des hypothèses de « segmentabilité » (atomistique) et de compositionnalité, en fondant la démarche inductive non pas sur les relations entre niveaux, mais sur *les principes d'organisation internes à chaque niveau de représentation*. L'hypothèse principale que je défends est, qu'à chacun de ces niveaux, le système formé par un ensemble d'unités tend à s'organiser d'abord à travers un ensemble régulier d'*alternances* et d'*oppositions*.

Cette idée trouve son origine première dans de nombreux travaux en morphologie (en particulier [Matthews, 1974, p. 185 et suivantes]), qui ont conduit à mettre en évidence l'organisation *paradigmatique* des systèmes flexionnels de nombreuses langues. Cette organisation se caractérise par un ensemble fixe et systématique d'oppositions fonctionnelles (par exemple *Singulier* vs. *Pluriel*), dont les marqueurs formels sont distribués, pour une classe flexionnelle donnée, en relation d'implication bi-univoque. Ces systèmes d'implications n'ont pas nécessairement de relation directe avec les catégories traditionnelles de la grammaire : ainsi en latin, la forme de la première personne du singulier de l'imparfait du subjonctif se "déduit" directement de celle de l'infinitif présent (exemple tiré de [Matthews, 1974, p. 194]), alors qu'il n'existe pas de relation logique entre la modalité exprimée par l'imparfait du subjonctif et celle exprimée par l'indicatif présent. Dans cette perspective, qui s'oppose à l'analyse morphématique classique, ce n'est pas tant la décomposition d'une forme en unités minimales qui compte, que son implication dans un réseau régulier d'oppositions formelles qui se corrèlent avec un ensemble d'oppositions fonctionnelles.

Sans reprendre à mon compte l'idée que la notion de paradigme pourrait servir de description de l'ensemble des données linguistiques[14], certains des principes que cette notion conduit à mettre en œuvre me semblent avoir un intérêt au delà de la description des systèmes flexionnels et sont susceptibles de s'appliquer à d'autres types de données. Cette position rejoint celle défendue par [Matthews, 1976, p. 267 et suivantes] en ce qui concerne l'analyse de données syntaxiques, et plus généralement par [Lavie, 2003]. Elle se trouve renforcée par les études sur corpus décrites dans [Lepage, 2004], qui fournissent une évaluation de la « densité paradigmatique » de corpus de phrases. En particulier, je défends, avec ces auteurs, l'idée que la notion de *paradigme*[15]

– permet, dans une large mesure, de se dispenser des hypothèses atomistes portant sur l'existence d'une segmentation en unités minimales et sur le fonctionnement compositionnel de ces unités. À l'inverse, elle

[14] En particulier, la notion de paradigme implique une systématicité dans la répartition des formes au sein d'un paradigme, systématicité qui est loin d'être avérée pour les données de morphologie dérivationnelle, *a fortiori* pour des données syntaxiques.

[15] Entendu ici dans un sens un peu plus précis que son sens habituel d'ensemble des formes possibles d'un lexème' : on suppose que ces ensembles de formes présentent des ressemblances dans leur organisation.

autorise à laisser les formes enregistrées dans la mémoire non analysées, potentiellement segmentables (réanalysables) de diverses manières en fonction de l'input avec lequel elles seront comparées.
– s'accomode d'un mécanisme de re-construction de la forme associée à une interprétation (dans le cas de la flexion, d'un ensemble de propriétés) qui se fonde sur le mécanisme de l'*analogie* : si *chiens*, comme pluriel de CHIEN, s'oppose au singulier *chien*, alors *chats* est le pluriel du lexème dont le singulier est *chat*.

Conclusion

La démarche présentée dans les sections qui suivent se fonde donc principalement sur la *réanalyse paradigmatique* d'un ensemble de données, c'est-à-dire la recherche, s'instantiant dans un contexte donné, d'oppositions *formelles* entre descriptions à un niveau de représentation qui soient corrélées *de manière stable* à des oppositions à un autre niveau de représentation. C'est dans ce contexte particulier que doit s'apprécier la conception et l'utilisation de méthodes à base d'analogies pour détecter et exploiter ces systèmes d'alternances formelles, méthodes qui sont développées dans les sections qui suivent. Cette perspective me rapproche de travaux récents conduits en linguistique [Itkonen and Haukioja, 1997, Lavie, 2003] sur les modélisations à base d'analogies.

En particulier, ce travail d'une part fournit un cadre possible pour l'apprentissage d'analogies structurelles et systémiques ; et d'autre part prolonge (en particulier à la section 2.4) les réflexions de Lepage sur le calcul d'analogies entre séquences sur un alphabet fini.

À l'inverse, il est important de souligner que, si le modèle d'inférence inductive développé ci-dessous s'apparente aux systèmes de raisonnement par cas, mes objectifs diffèrent singulièrement des travaux conduits en Intelligence Artificielle sur le raisonnement par analogie (cf. la section 2.1.1), qui visent à fournir des modèles *généraux* et *cognitivement réalistes* de résolution de problèmes par analogie.

2.2 Un modèle d'inférence analogique

Cette section est consacrée à une présentation du modèle d'inférence analogique initialement proposé dans [Pirrelli and Yvon, 1999a]. Dans ce modèle, les données linguistiques sont représentées par des structures attribut-valeur et sont supposées avoir été intégralement mémorisées lors d'une étape d'apprentissage préalable. L'inférence analogique consiste à généraliser l'ensemble d'apprentissage de la manière suivante : pour un nouvel exemple présenté au système, des relations de proportionnalité partielles sont dans un premier temps identifiées, qui permettent dans un second temps la reconstruction analogique des attributs inconnus. Les différents composants de ce modèle sont successivement présentés dans les paragraphes qui suivent : la notion de rapport de proportion ; le problème de l'identification des rapports de proportion existant dans une collection d'objets ; enfin la question du transfert analogique.

2.2.1 Rapports de proportion

Dans cette première version du modèle, on considère que les représentations linguistiques manipulées prennent la forme de structures attribut-valeur ou structures de traits, les attributs prenant soit des valeurs numériques, soit des valeurs symboliques dans un ensemble fini, soit des valeurs dans des ensembles de mots (séquences finies sur un alphabet fini). Ces structures de traits sont donc, par construction, non récursives. Pour chacun de ces types d'attribut, la notion de proportion est définie comme suit.

Définition 2.1. *Soient x_1, x_2, x_3 et x_4 quatre variables. Un rapport de proportionnalité, noté $x_1 : x_2 :: x_3 : x_4$ lie ces variables :*

(i) si x_1, x_2, x_3 et x_4 sont numériques (dans $\mathbb{R}$ ou dans $\mathbb{N}$) et

$$\frac{x_1}{x_2} = \frac{x_3}{x_4}$$

.

(ii) si x_1, x_2, x_3 et x_4 prennent des valeurs dans un ensemble fini et $(x_1 = x_2 \wedge x_3 = x_4)$ ou bien si $(x_1 = x_3 \wedge x_2 = x_4)$.

(iii) si x_1, x_2, x_3 et x_4 sont des mots de $\Sigma^\star$, et s'il existe quatre mots de $\Sigma^\star$, t, u, v, w tels que :

$$\begin{cases} \text{soit} & x_1 = uv & x_2 = uw & x_3 = tv & x_4 = tw \\ \text{soit} & x_1 = uv & x_2 = tv & x_3 = uw & x_4 = tw \end{cases}$$

La définition (i) est simplement la définition des rapports de proportionnalité sur $\mathbb{R}$ et $\mathbb{N}$. Le cas (ii) restreint les analogies impliquant des symboles aux configurations $a : a :: b : b$ et $a : b :: b : a$ (cf. les propriétés usuelles de l'analogie introduites au début de la section 2.1.1). On peut l'illustrer par des relations de proportionnalité entre parties du discours : Nom : Nom :: Verbe : Verbe ou encore Adj : Verbe :: Adj : Verbe. La définition (iii) concerne un troisième type de représentation usuel en linguistique : les séquences ordonnées sur un alphabet fini. Conformément à l'intuition, elle exprime que la relation de proportionnalité correspond à des alternances de préfixes et de suffixes. Celle définition est illustrée par le quadruplet de formes orthographiques (*jouer*, *joueur*, *manger*, *mangeur*) définit une proportion analogique, avec u =*jou*, v =*er*, t =*mang* et w =*eur*. Cette version de l'analogie entre séquences est utilisée pour les expérimentations présentées aux sections 2.3.2 et 2.3.1. Ce n'est toutefois pas la seule possible (voir par exemple [Lepage, 1998]) et la section 2.4 en propose une version plus à la fois plus générale et susceptible de s'étendre à une conception graduelle de l'analogie entre mots.

De ces définitions, on étend la notion de proportionnalité à des structures de traits non-récursives[16].

Définition 2.2. *Quatre structures de traits définissent une proportion analogique si et seulement si :*
– *elles contiennent les même traits ;*
– *les valeurs de chacun des traits (numérique, symbolique, séquences) définissent terme à terme des proportions analogiques ;*

La figure 2.2 illustre cette notion pour des structures attribut-valeur simples :

$$\begin{bmatrix} \text{graphie} : \textit{joueur} \\ \text{cat} : \text{Nom} \\ \text{genre} : \text{masc} \\ \text{nombre} : \text{sing} \\ \text{pers} : \text{-} \end{bmatrix} : \begin{bmatrix} \text{graphie} : \textit{penseur} \\ \text{cat} : \text{Nom} \\ \text{genre} : \text{masc} \\ \text{nombre} : \text{sing} \\ \text{pers} : \text{-} \end{bmatrix} :: \begin{bmatrix} \text{graphie} : \textit{jouent} \\ \text{cat} : \text{Verbe} \\ \text{genre} : \text{-} \\ \text{nombre} : \text{plur} \\ \text{pers} : 3 \end{bmatrix} : \begin{bmatrix} \text{graphie} : \textit{pensent} \\ \text{cat} : \text{Verbe} \\ \text{genre} : \text{-} \\ \text{nombre} : \text{plur} \\ \text{pers} : 3 \end{bmatrix}$$

FIG. 2.2 – Rapport de proportion entre structures de traits

Les analogies entre structures, qui établissent simultanément des rapports de proportionnalité entre différents niveaux de la description d'un objet linguistique, sont appelées *analogies multi-dimensionnelles*.

2.2.2 Transfert analogique

Le transfert analogique consiste à inférer ou plutôt reconstruire le quatrième terme d'une proportion analogique à partir des trois autres : il s'agit en fait de résoudre une équation de type : $x_1 : x_2 :: x_3 : x$, où x est

[16] Cette définition sera généralisée au cas des structures de traits arbitraires à la section 2.4.3.

la valeur inconnue. Le résolution de ce type d'équations dépend, à nouveau, du type des attributs impliqués. S'ils prennent des valeurs numériques, la résolution est immédiate.

S'ils prennent des valeurs symboliques, il faut considérer trois cas de figures pour résoudre $x_1 : x_2 :: x_3 : x$:
– si $x_1 = x_2$ alors $x = x_3$
– si $x_2 = x_3$ alors $x = x_2$
– sinon l'équation n'a pas de solution et x est indéfini.
Si, enfin, il s'agit de séquences de symboles, il faut également distinguer trois cas :
– s'il existe t, u, v, w tels que $x_1 = uv, x_2 = tv, x_3 = uw$, alors $x = tw$
– s'il existe t, u, v, w tels que $x_1 = uv, x_2 = uw, x_3 = tv$, alors $x = tw$
– sinon, x n'est pas défini
Notons que pour les mots, la solution d'une équation analogique n'est pas nécessairement unique, les deux premières conditions n'étant pas mutuellement exclusives. Ainsi, $a : ba :: ca : x$ possède deux solutions : $x = bca$ et $x = cba$.

La notion de transfert analogique s'étend à des vecteurs (structures) d'attributs : la solution d'une équation analogique entre structures correspond à une structure dont chacun des attributs a pour valeur la solution de l'équation analogique portant sur cet attribut. Si l'une de ces équations n'a pas de solution, alors il n'existe pas de structure solution.

2.2.3 Recherche et comparaison de motifs analogiques

Le problème abordé dans cette section est celui de la généralisation analogique d'une base d'exemples représentés par des structures attribut-valeur. Face à un nouvel objet u représenté par le vecteur d'attribut $\mathbf{x}$, dont seuls quelques composantes sont connues, il s'agit d'inférer, à partir des instances connues dans leur intégralité, les attributs qui manquent. [Federici et al., 1996, Yvon, 1996, Pirrelli and Yvon, 1999b] proposent diverses solutions à ce problème, qui impliquent les deux étapes suivantes :

– La recherche de toutes les proportions analogiques dans lesquelles entrent (la sous-structure correspondant aux attributs connus de) $\mathbf{x}$ et trois objets connus. Cette phase de recherche est la plus exigeante en temps de calcul, puisqu'elle demande de vérifier, pour chaque $\mathbf{u}^{(i)}$ connu, s'il existe dans la base de données $\mathbf{u}^{(j)}$ et $\mathbf{u}^{(k)}$ tels que $\mathbf{x} : \mathbf{x}^{(\mathbf{i})} :: \mathbf{x}^{(\mathbf{j})} : \mathbf{x}^{(\mathbf{k})}$, la proportion étant définie par rapport aux composantes connues de $\mathbf{u}$ qui spécifient $\mathbf{x}$. L'ensemble des candidats possibles pour $\mathbf{x}^{(\mathbf{j})}$ et $\mathbf{x}^{(\mathbf{k})}$ se déduit directement des valeurs des attributs de $\mathbf{x}$ et $\mathbf{x}^{(\mathbf{i})}$. [Pirrelli and Yvon, 1999b] montre que, pour les tâches que nous avons étudiées, cet ensemble de candidats reste d'une taille suffisamment faible pour être exploré exhaustivement.
– Pour chaque triplet extrait, il s'agit ensuite de calculer le résultat de l'équation analogique pour la sous-structure correspondant aux attributs inconnus. Lorsque l'équation a une solution, elle est ajoutée à l'ensemble des hypothèses.
Une procédure de sélection de la meilleure structure candidate est ensuite mise en œuvre, qui peut intégrer différents types de préférences : par exemple, choisir la solution qui est attestée par le plus grand nombre de triplets, comme dans les expérimentations décrites dans [Pirrelli and Yvon, 1999a, Yvon, 1999] ; ou en utilisant un critère fondé sur la longueur du *suffixe commun* entre la chaîne cible et la chaîne source [Federici et al., 1996]. Suivant les domaines, cette procédure de sélection peut être adaptée, en fonction des connaissances disponibles.

Ces deux étapes sont représentées sur la figure 2.3 dans le cas d'une tâche de prononciation de mots inconnus (voir également ci-après à la section 2.3.1).

En résumé, l'inférence analogique désigne donc une famille de méthodes qui sont complètement spécifiées par la définition de procédures permettant de (a) vérifier et résoudre des équations analogiques, et (b) comparer des solutions lorsque plusieurs solutions alternatives sont en compétition.

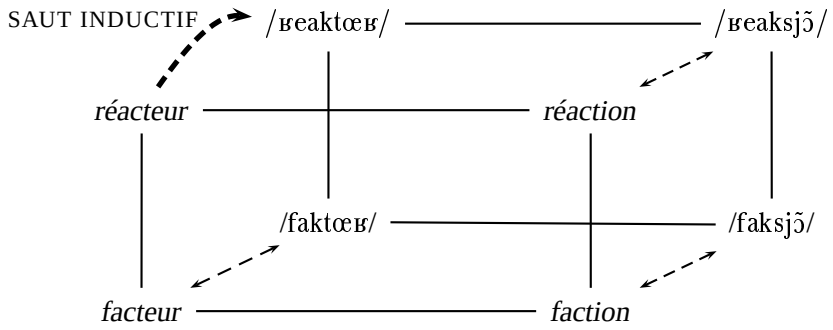

FIG. 2.3 – Transfert analogique pour la tâche de prononciation

Pour le mot inconnu 'réacteur' présenté sous une forme orthographique, la phase de recherche permet d'exhiber le rapport de proportion 'réacteur : réaction :: facteur : faction'. En résolvant l'équation analogique : '/faksjɔ̃/ : /faktœʁ/ :: /ʁeaksjɔ̃/ : ?', on déduit la prononciation : /ʁeaktœʁ/.

Il importe dès à présent de souligner ici deux propriétés importantes de cet apprenti. En premier lieu, il généralise la notion de catégorisation, qui correspond au cas particulier où seule une composante symbolique du vecteur de sortie est prédite. En second lieu, il maintient une complète symétrie entre entrée (ce qui est connu de l'objet à analyser) et sortie (ce qui ne l'est pas). En fait, dès mon travail de thèse [Yvon, 1996], j'ai présenté des tests sur la génération de formes orthographiques à partir de formes phonétiques ; cette capacité est de nouveau exploitée à la section 2.3.2, pour une tâche de génération de formes fléchies.

2.3 Quelques expérimentations

Cette section décrit quelques-unes des expérimentations conduites avec le modèle d'inférence analogique présenté à la section précédente : elles portent sur la tâche de conversion graphème-phonème (section 2.3.1), puis sur la tâche d'analyse morphologique (section 2.3.2). La discussion critique qui la clôt tire quelques leçons de ces expérimentations et annonce par avance les évolutions présentées dans la section 2.4.

2.3.1 Prononciation de mots inconnus

En toile de fond

La conversion graphème-phonème, vue comme une (sous-)tâche particulière du traitement automatique des langues, est introduite à la section A.0.7 ; diverses simplifications usuellement effectuées pour pouvoir appliquer des techniques d'apprentissage automatique à cette tâche y sont également discutées.

Ces travaux sur la prononciation de mots inconnus par recherche et reconstruction de rapports de proportion entre mots existant dans des lexiques, poursuivent les développements initiés dans [Yvon, 1996, Chap. 4]. Ces travaux utilisent des représentations dans lesquelles le seul attribut disponible en entrée est la forme orthographique et le seul attribut calculé en sortie est la séquence d'unités phonémiques représentant la prononciation. Ces travaux, qui comparent extensivement l'apprentissage par analogie avec d'autres apprentis ont conduit aux deux conclusions suivantes :

– la construction de rapports de proportion entre formes orthographiques (tels que ceux présentés dans la section précédente) est une méthode « sûre » pour construire la prononciation de mots inconnus. Lorsque de tels rapports existent dans le lexique, ils permettent de produire une prononciation plus précise que des

méthodes utilisant des algorithmes d'apprentissage traditionnels, avec le bénéfice supplémentaire qu'ils dispensent de l'étape préalable d'alignement du lexique d'apprentissage.
– si la précision est très bonne, le rappel est en revanche insuffisant : dans la mesure où l'on manipule des analogies sur de séquences de symboles, l'étape de recherche comme l'étape de transfert (résolution d'équations) peuvent échouer, conduisant à une échec global de la procédure.

Développements

Les développements sur ce thème ont principalement porté sur l'utilisation d'analogies multi-dimensionnelles, avec pour ambition d'apprendre, en utilisant des codages plus riches des entrées, des systèmes de prononciation plus réalistes, ie. répondant de manière plus complète aux besoins applicatifs réels (cf. la discussion de la section A.0.7). En particulier, [Yvon, 1999] montre comment il est possible, en utilisant des analogies multi-dimensionnelles :

1. de calculer *simultanément* la chaîne segmentale, la structure syllabique et le motif accentuel pour une chaîne orthographique en entrée ;
2. d'utiliser des descripteurs supplémentaires (eg. des indications de catégories morpho-syntaxiques), qui permettent effectivement de désambiguïser la prononciation pour une forme ambiguë (homographes-hétérophones) ;
3. d'utiliser des classes de symboles dans des représentations multi-dimensionnelles. Dans ces tests, par exemple, l'introduction des classes 'consonne' et 'voyelle' permet de représenter la graphie (ou l'orthographe) d'un mot sur deux flux séparés : d'un côté, la séquence vocalique, de l'autre, la séquence consonantique. Avec une telle représentation, l'entrée *care* est recodée en une paire (*CaCe*,cVrV).

   Ce changement de représentation vise à opérer des analogies dans des espaces de représentation plus « denses », avec pour objectif d'augmenter le rappel.

La série de tests correspondant au point (1) révèle, de manière un peu paradoxale, que complexifier la sortie du système de transcription permet d'en améliorer le comportement global. Il apparaît en fait que les différentes dimensions de la sortie se contraignent mutuellement, augmentant la précision globale du système, sans dégrader significativement le rappel.

Les expérimentations correspondant à (2) montrent que l'introduction d'un attribut supplémentaire en entrée permet également d'améliorer la précision du transcripteur, en particulier en ce qui concerne les homographes hétérophones, ici encore pratiquement sans impact sur le rappel.

Les expérimentations concernant le point (3) sont moins concluantes : si le rappel augmente, le changement de représentation envisagé conduit à une réduction très significative de la précision ; il s'avère toutefois trop fruste pour maintenir la précision à un niveau raisonnable. Il s'avère en fait que remplacer toutes les consonnes par un symbole unique complexifie énormément la prédiction des voyelles phonétiques.

Dernier résultat intéressant de cette série d'expérimentations : la sortie correcte est bien plus souvent construite qu'elle n'est proposée comme première solution. Ceci suggère qu'optimiser la stratégie de reclassement des sorties est une direction possible d'amélioration des résultats.

Un bilan

Ces travaux ont donc permis de mettre en œuvre un apprenti capable de traiter, sans changement de représentation, une version plus réaliste de la tâche de conversion graphème-phonème, dans laquelle les entrées comme les sorties sont multi-dimensionnelles. En plus du problème du silence, cette modélisation reste insatisfaisante sur un aspect : la production de variantes de prononciation. Cette extension requiert la capacité

de traiter des attributs dont la valeur est non plus une prononciation mais un ensemble (fini) de prononciations possibles, et donc de savoir résoudre des équations analogiques entre des ensembles finis de chaînes. L'extension du modèle d'analogie entre séquences à des langages finis proposée à la section 2.4 fournit un cadre dans lequel une telle généralisation deviendra possible.

2.3.2 Analyse morphologique de mots inconnus

En toile de fond

Le terme *d'analyse morphologique* recouvre en traitement automatique des langues des réalités diverses, qui sont discutées à la section A.0.7. Ce sous-domaine constitue un champ d'expérimentation privilégié pour les méthodes analogiques, dans la mesure où le concept de *paradigme*, au sein desquels s'expriment exemplairement des rapports de proportion, a été élaboré et utilisé précisément dans le but de décrire les régularités morphologiques (en particulier de la flexion des langues indo-européennes).

Les travaux décrits dans cette section portent d'une part sur la génération automatique de formes fléchies à partir d'un lemme et de caractérisques morpho-syntaxiques, d'autre part sur l'identification de relations dérivationnelles dans un lexique, ayant pour but l'identification de la racine d'une forme ou d'un lexème. Dans les deux cas, je me suis limité à la manipulation de représentations linéaires « plates » pour les formes, limitation imputable à la modélisation des analogies entre séquences utilisée dans ces travaux.

Génération (et analyse) de formes fléchies

La tâche envisagée consiste à prédire une forme fléchie à partir d'une représentation d'un lemme, et des traits morpho-syntaxiques de la forme à atteindre ; ou inversement à analyser une forme inconnue sous la forme d'un lemme et des traits morpho-syntaxiques associés.

Pendant ma thèse [Yvon, 1996], j'ai proposé une première modélisation très simple des systèmes de conjugaison, en utilisant un encodage linéaire de l'entrée. Pour la tâche de génération de formes verbales de l'anglais, problème classique de l'apprentissage automatique depuis [Rumelhart and McClelland, 1986], l'entrée du système se présente sous la forme d'une séquence *LEMME+SLOT*, où *SLOT* peut prendre une des cinq valeurs suivantes : *B* pour la forme présent (hors 3ème personne), *3* pour le présent (3ème personne), *P* pour la forme passé, *D* pour le passé composé, et *G* pour le gérondif. La sortie est également mono-dimensionnelle, et correspond à une représentation (orthographique ou phonétique) de la forme fléchie correspondante. Le lexique est donc intégralement représenté comme une liste de telles paires, par exemple : (*GIVE+B, give*), (*GIVE+P, gave*), etc. Les résultats obtenus par application d'analogies unidimensionnelles se comparent très favorablement à ceux obtenus, sur ce même problème, par d'autres techniques d'apprentissage automatique (voir, sur ce problème, en particulier [Rumelhart and McClelland, 1986, Pinker and Prince, 1988, Lachter and Bever, 1988, MacWhinney and Leinbach, 1991, Ling and Marinov, 1993]).

Ces expérimentations préliminaires sur l'apprentissage de systèmes flexionnels ont donné suite à de nouveaux travaux, utilisant des représentations plus « naturelles » des entrées et sorties, sous la forme de structures multi-dimensionnelles telles que celles prônées par les partisans d'une morphologie lexématique [Fradin, 2003]. Une entrée typique pour la tâche de génération sera alors une représentation multi-dimensionnelle comme celle représentée à la figure 2.4, la sortie désirée étant la forme graphique (ou phonétique) associée à cette représentation.

J'ai utilisé le modèle décrit à la section 2.2 pour des tâches d'analyse et de génération de formes, pour diverses langues (en particulier le français et l'anglais). Les résultats (non publiés) obtenus pour une telle

$$\begin{bmatrix} \text{lexeme} : \text{MANGER} \\ \text{cat} : \text{V} \\ \text{temps} : \textit{pres} \\ \text{mode} : \textit{ind} \\ \text{pers} : \textit{3eme} \\ \text{nombre} : \textit{sing} \end{bmatrix}$$

FIG. 2.4 – Représentation multi-dimensionnelle de la forme *mange*

tâche confirment la pertinence de la modélisation par analogie pour apprendre les régularités des systèmes flexionnels de ces langues.

La table 2.1 présente ainsi les résultats de l'analyse de formes anglaises à partir de sous-échantillons aléatoires du lexique de formes CELEX [Burnage, 1990] : dans cette série de tests, 50 000 formes[17] sont aléatoirement (avec probabilité uniforme, sans remise) tirées dans le lexique complet puis aléatoirement réparties entre données d'apprentissage (90%) et données de test (10%) ; l'entrée est une forme graphique et la sortie est restreinte aux deux attributs (lemme, catégorie syntaxique). On est donc proche ici de la tâche « divination » morpho-syntaxique (définie à la section A.0.7), bien que les tests envisagés dans ce travail portent en réalité sur des formes du lexique CELEX et pas sur de « vraies » formes inconnues.

cat.	Rappel	Précision (LEMME)	Précision (CAT)	Précision (L+C)
Nom	50.1	58.6	27.6	24 .1
Vrb	96.5	90.9	86.7	85.5
Adj	56.7	37.3	22.3	20.0
Total	73.4	53.7	65.4	56.0

TAB. 2.1 – Analyse de formes

Pour les noms, environ 50% des formes sont analysées analogiquement, avec une précision pour la prédiction de l'attribut LEMME *de 58.6%, et de 27.6% pour la prédiction de l'attribut* CATÉGORIE.

La table 2.2 donne les résultats, détaillés par catégorie syntaxique majeure, de la tâche inverse (la génération) : le lemme et l'ensemble des traits étant connus, il s'agit de calculer la forme correspondante : le cadre expérimental est donc tout à fait semblable à celui utilisé dans [Lepage, 1999b], qui se limite toutefois aux flexions verbales. Cette expérimentation a été conduite sur le français, en échantillonnant des données d'apprentissage et de test dans le lexique Multext-FR [Ide and Veronis, 1994], avec un protocole identique à celui utilisé pour les expérimentations qui concernent l'anglais.

cat.	précision	% calculée
Nom	79.8	97.2
Vrb.	90.4	99.7
Adj.	53.9	99.2
Total	78.3	99.5

TAB. 2.2 – Génération de formes

Pour 79.8% des noms, la forme proposée comme meilleure solution est la forme correcte, dans 97.2% des cas, la forme correcte est produite (mais pas nécessairement choisie comme étant la meilleure). Dans cette série d'expérimentations, le rappel est de 100%.

Les résultats préliminaires présentés ci-dessus présentent les mêmes traits que ceux présentés pour la tâche

[17] Compte-tenu de la petite taille de l'ensemble d'apprentissage, il est raisonnable de penser que ces résultats sous-estiment les performances réelles de ce générateur à base d'exemples.

de prononciation : la précision moyenne est bonne, voire très bonne si l'on considère le nombre de cas où la solution correcte est construite ; en revanche, et c'est surtout net pour les tâches d'analyse, pour lesquels l'espace d'entrée est moins densément peuplé, le rappel reste insuffisant. Il est intéressant de constater, même si cette discussion ne sera pas plus développée dans le cadre de ce document, que la comparaison des résultats obtenus sur plusieurs langues, ou, pour des données d'une langue, entre les différentes catégories de lexèmes, est susceptible d'apporter des éclairages sur la structure de la langue, en mettant différentiellement en évidence les zones à forte ou faible densité paradigmatique.

Identification de chaînes dérivationnelles

L'évaluation de la portée des méthodes à base d'analogie nous a également conduit à considérer une tâche d'analyse morphologique dérivationnelle. [Pirrelli and Yvon, 1999b] présente une série d'expérimentations conduites sur le lexique anglais de la base de données CELEX [Burnage, 1990]. Il s'agit de prédire deux attributs d'un mot inconnu à partir de sa forme orthographique : sa catégorie morpho-syntaxique principale (en douze classes) et la base (ou la racine) associée au mot. Le contexte est donc très proche de la tâche d'analyse de formes présentées ci-dessus, à ceci près qu'ici on analyse des lemmes, alors que précédemment on analysait des formes.

Les principaux enseignements de ces expérimentations sont les suivants :
– les performances atteintes par l'apprenti analogique sont très satisfaisantes, en particulier pour les lemmes qui sont véritablement construits morphologiquement (par opposition aux mots composés, aux emprunts...) : pour cette classe de lemmes, le rappel est de l'ordre de 98%, la précision de la prévision de la partie du discours de 89% et la précision atteinte en termes d'identification de la base dépasse 87%.
– les performances obtenues sur la tâche d'identification de la base témoignent la densité analogique du lexique et valident l'hypothèse selon laquelle, du point de vue des modifications de surface, les phénomènes dérivationnels, comme les phénomènes flexionnels, peuvent être décrits comme s'organisant en quasi-paradigmes.
– les phénomènes de *dérivation zéro*, qui correspondent à un simple changement de catégorie *sans changement de forme*[18] obscurcissent la structure globalement paradigmatique du lexique utilisé et contribuent à dégrader sensiblement les performances.

Divers prolongements de ces travaux sont envisageables, en particulier dans la direction d'enrichir l'entrée du système d'inférence par de nouvelles dimensions, correspondant, par exemple, à une description statistique du voisinage lexical dans lequel un lemme non répertorié apparaît, à la manière des modèles développés dans [Schone and Jurafsky, 2000, 2001] ; ou encore des informations sémantiques tirées de ressources lexicales telles que Wordnet, à l'instar du travail présenté dans [Hathout, 2001].

Ces prolongements n'ont pas été envisagés, dans la mesure où ils demanderaient en préalable :
– la définition plus précise d'un contexte applicatif dans le cadre duquel une réelle évaluation des différents modèles puisse être conduite, ce qui était loin d'être le cas des expérimentations décrites dans [Pirrelli and Yvon, 1999a] : par exemple une tâche de « divination » de la catégorie et de traits morpho-syntaxiques de mots inconnus ou encore d'analyse de lemmes hors lexique.
– une meilleure appréhension des limites de la démarche à base d'analogie, conduisant à l'élaboration de stratégies hybrides. Il existe en effet différents mécanismes à l'origine de l'apparition de mots inconnus dans les textes : en plus des dérivations régulières et productives, pour l'analyse desquels l'inférence analogique est indéniablement performante, il existe des phénomènes d'emprunts, de composition, etc. qui ne sont pas tous également bien capturés par le modèle.

[18] Ainsi, en français de nombreux adjectifs sont-ils aussi enregistrés comme formes nominales ; en anglais on observe de nombreuses paires (Nom, Verbe) tels que *a record-to record* ; *a program - to program* etc.

Conclusion

Ces diverses expérimentations sur des tâches d'analyse morphologique ont un double mérite : d'une part le développement de systèmes exhibant des performances proches de l'état de l'art ; d'autre part, la mise en place d'un modèle[19] d'apprentissage de la morphologie s'affranchissant de la notion de morphème et donc plus conforme aux descriptions linguistiques modernes d'une morphologie lexématique (voir e.g. [Anderson, 1992, Aronoff, 1994, Fradin, 2003]). À l'aide de ce modèle, nous avons pu valider l'hypothèse d'une organisation essentiellement paradigmatique des lexiques, y compris des lexiques de lemmes.

Les problèmes rencontrés sont sensiblement les mêmes que pour la conversion graphème-phonème, à savoir (i) un rappel trop faible, (ii) des durées de traitement trop longues et (iii) une incapacité à sélectionner la bonne forme parmi les hypothèses proposées.

2.3.3 Discussion

Cette section vise à tirer quelques leçons des différents tests portant sur les tâches d'acquisition de connaissances lexicales, en montrant comment le modèle d'apprentissage par analogie apporte des réponses effectives à certaines des limitations des modèles classificatoires en traitement automatique des langues discutées à la section 2.1.4.

J'ai déjà eu l'occasion de souligner l'intérêt, pour de nombreuses tâches du traitement automatique des langues, de stratégies d'apprentissage paresseux, qui offrent une prise en charge directe des nombreuses idiosyncrasies (par apprentissage par cœur) qu'on peut rencontrer. J'ai également mentionné que ces stratégies présentent l'avantage que le saut inductif est toujours opéré dans un contexte particulier, une fois connue la nouvelle instance à analyser.

Cette section évoque d'autres bénéfices de cette approche, en discutant successivement les questions du silence, des représentations linguistiques et de la localité, en reprenant, dans ses grandes lignes, une discussion conduite dans [Pirrelli and Yvon, 1999a].

Le silence : un contre, deux pour

Le mécanisme d'apprentissage par analogie utilisé dans ces expérimentations souffre d'une faiblesse évidente : il est dans l'incapacité de proposer une sortie pour chacune des entrées. Dit en d'autres termes, l'apprentissage par analogie, s'il a une excellente *précision* sur les tâches considérées ci-dessus, souffre d'un *rappel* trop faible.

Deux situations d'échec sont observées : soit il est impossible de construire des quadruplets analogiques dans l'espace d'entrée, signalant que la cible est trop différente des exemples rencontrés lors de l'apprentissage ; soit les quadruplets retrouvés ne permettent pas de construire une solution dans l'espace de sortie, ce qui indique l'anomalie de ces proportions analogiques.

Ce faible rappel se mesure aisément pour les tâches d'enrichissement lexical : disposant d'un vocabulaire de taille $|V|$, le système ne pourra produire des sorties que pour les mots qu'il est possible de reconstruire à partir de 3 mots de V, soit pour au plus $0(|V|^3)$ nouvelles entrées. Ce conservatisme est naturellement amplifié lorsque l'on envisage des analogies portant sur des structures multi-dimensionnelles.

Cette caractéristique de l'apprentissage par analogie contraste négativement avec la capacité de généralisation des apprentis statistiques, qui sont capables de produire une sortie pour toute nouvelle entrée. Cette lacune est indiscutablement pénalisante pour un certain nombre d'applications. Elle paraît également rédhi-

[19] Le seul modèle ? voir également [Neuvel and Fulop, 2002].

bitoire pour envisager d'appliquer directement ce modèle à des tâches d'analyses syntaxiques, dans lesquelles l'organisation paradigmatique des données d'apprentissage est bien moins évidente que pour des tâches lexicales[20]. Nuançons un peu cette observation en soulignant que :

– Il est possible, dans une certaine mesure, de contrôler le niveau de rappel en augmentant les données d'apprentissage, ou encore en affaiblissant les contraintes sur les analogies considérées comme valides. Les travaux qui seront présentés à la section 2.4 constituent une avancée permettant, pour partie, de répondre à ce problème de rappel.
– Les entrées sur lesquelles l'apprentissage par analogie échoue sont loin d'être distribuées aléatoirement. Bien au contraire, elles correspondent le plus souvent, pour des tâches lexicales, à des formes graphiques que l'on pouvait qualifier de déviantes : emprunts, dérivés irréguliers, etc. Ceci permet d'apprécier plus positivement ce faible rappel, qui témoigne d'une capacité à identifier les entrées qui sont susceptibles de poser problème à tout système de traitement automatique. Dans certains contextes, cette capacité peut être vue comme un atout. Ainsi, par exemple, pour des tâches de constitution semi-automatique de ressources, dans lesquelles il faut pouvoir signaler automatiquement à l'opérateur humain les entrées problématiques ; ou encore pour mettre en œuvre des stratégies d'apprentissage par *bootstrapping* [Abney, 2002] (voir également la section 1.3.1) ou encore d'apprentissage actif.

Les représentations

La mise en œuvre d'outils de classification supervisée pour résoudre des tâches d'analyse linguistique présuppose l'adoption d'une représentation des données d'apprentissage dans lesquelles chaque entrée est décrite par un vecteur de caractéristiques symboliques ou numériques. J'ai déjà mentionné (à la section 2.1.4), un des bénéfices de l'apprentissage par analogie, qui permet d'effectuer un traitement « holistique » sur les représentations linguistiques correspondant à des séquences et de se dispenser de la segmentation de l'entrée en unités atomiques et de leur appariemment préalable avec les éléments de la sortie. Cette section vise à pointer un second avantage de cette méthodologie, à savoir le fait qu'elle permet de travailler sur des représentations qui me semblent plus *naturelles*.

Les représentations linguistiques sont, de manière inhérente, multi-dimensionnelles, c'est-à-dire qu'elles enregistrent des informations qui portent sur différents niveaux de représentation. La projection de l'ensemble de ces informations dans un unique vecteur de caractéristiques conduit à faire disparaître cette multi-dimensionnalité et à construire des représentations qui sont inutilement redondantes et à introduire de manière artificielle des corrélations entre les attributs utilisés dans la description.

Considérons ainsi de nouveau la tâche de conversion graphème-phonème. Il est connu que de multiples facteurs aident à déterminer la prononciation d'une chaîne orthographique : suivant les langues, la connaissance de la position des accents lexicaux, de la structure syllabique, de l'histoire dérivationnelle, des traits morpho-syntaxiques, voire de la catégorie syntaxique peuvent jouer un rôle. Dans un modèle de classification, dans lequel chaque forme est représentée comme une séquence de vecteurs d'attributs (un vecteur pour chaque position orthographique), la prise en compte de ces informations conduit à la construction de représentations dans lesquelles, par exemple, l'information de catégorie, qui est constante pour une forme donnée, sera dupliquée pour chacune des lettres. De même, le marquage des positions accentuées, qui par définition ne concerne réellement que les voyelles, devra être spécifié y compris pour les consonnes.

Cette « déformation » des représentations linguistiques a un double prix : elle complique inutilement l'étape de prétraitement des entrées d'apprentissage ; plus sérieusement, en augmentant de manière injustifiée la taille de l'espace de représentation, elle a un effet automatique de dispersion des données qui rend plus difficile la détection (et l'expression) des régularités. Ainsi, pour prolonger l'exemple évoqué ci-dessus de la conversion graphème-phonème, l'introduction d'un attribut catégoriel dans les descriptions des lettres

[20] Voir toutefois [Lepage, 2004], qui fournit des éléments quantitatifs permettant de relativiser ce constat.

va demander que les règles de prononciation pour les graphèmes non-flexionnels soient réapprises pour chacune des catégories, alors pourtant qu'elles valent de façon transcatégorielle.

Le travail présenté à la section 2.4.3 vise à accentuer encore cet avantage des approches par analogie, dans la mesure où il étend le registre des structures sur lesquelles peuvent être définies des analogies, permettant ainsi de travailler directement avec toutes les représentations utilisées par les linguistes (arbres, langages finis, structures de traits, etc.).

La localité

Un autre trait des tâches d'analyse linguistique est que l'influence des divers éléments de contexte s'opère à une distance variable. Pour autant, les principales influences sont tenues pour être accessibles localement, permettant, de se limiter à une fenêtre de contexte syntagmatique de taille bornée. Ainsi, la détermination de l'étiquette catégorielle d'une forme ne se fonde-t-elle, dans la majorité des systèmes, que sur les formes proches (deux ou trois formes adjacentes), au détriment des influences distantes, qui sont beaucoup plus difficiles à capturer. Ce point est largement documenté et je discute ici un seul exemple, toujours tiré de mes expériences pour la tâche de conversion graphème-phonème.

En anglais, la règle de "relâchement tri-syllabique" (*tri-syllabic laxing*) donne lieu, du point de vue de la prononciation, à des alternances *régulières* entre une voyelle tendue et une voyelle lâche dans des formes apparentées dérivationnellement. C'est le cas, par exemple, de la paire *hostile*/*hostility* : dans le premier, le '*i*' donne lieu à un voyelle tendue (/aj/), alors que dans le second, après dérivation, on enregistre une réalisation relâchée (/ɪ/). L'apprentissage, par un classifieur, de cette alternance demande de détecter l'influence conjointe d'un '*i*', d'un '*t*' puis d'un '*y*' situés respectivement deux, trois et quatres positions après le '*i*' à prononcer. Cette détection doit, de surcroît, distinguer ce contexte particulier de toutes les autres configurations, extrêmement majoritaires, dans lesquelles ces trois positions n'ont pas d'influence sur la prononciation de '*i*'. Il s'agit ici d'un cas où la rareté des données d'apprentissage résulte en fait d'un choix de représentation inapproprié. À l'inverse, l'approche par analogie, qui permet de faire l'économie de l'hypothèse de compositionalité, permet d'identifier correctement de telles « micro-régularités » et de prédire la bonne prononciation.

Cet exemple illustre le fait que la rareté des données d'apprentissage, évoquée comme une des difficultés majeures pour la modélisation statistique des langues, ne se pose que relativement à un choix de représentation particulier, qui s'appuie sur une hypothèse de compositionalité. Ces approches sont donc confrontées au dilemme suivant : soit prendre en compte peu de descripteurs, conduisant à négliger des influences non-locales, ou des influences qui ne s'exercent que pour une sous-population des données ; soit intégrer de multiples descripteurs, au risque de ne pouvoir estimer de manière fiable leurs interactions par manque de données.

2.3.4 Conclusion

Dans cette section, j'ai présenté les résultats d'une série d'expérimentations d'un algorithme d'apprentissage par analogie, portant sur diverses tâches d'analyse lexicale. Cette approche conduit à des performances proches de l'état de l'art, en particulier, elle permet d'obtenir une très bonne précision. Il existe d'autres bénéfices à l'approche par analogie, qui l'opposent aux approches fondées sur l'utilisation de classifieurs. Il apparaît toutefois que la notion d'analogie entre séquences sur un alphabet fini telle que définie à la section 2.2 présente un certain nombre de faiblesses.

D'un point de vue pratique, sa rigidité conduit à l'échec de la procédure d'inférence, par manque de triplets candidats, ou par échec du transfert dès lors que l'input est trop différent des exemples disponibles. Ces échecs font que les résultats d'un système analogique ne sont pas toujours directement comparables à ceux

d'autres systèmes exploitant des méthodes d'apprentissage ; ils limitent également l'applicabilité opérationnelle de ces systèmes.

À l'inverse, lorsque l'entrée du système est trop proche des données d'apprentissage, l'étape de recherche de triplets analogiques donne lieu à un nombre inutilement grand de triplets et est source d'inefficacité. L'amélioration de cette procédure passe donc également par la conception de méthodes permettant de limiter le nombre d'analogues potentiels aux seuls exemples réellement utiles.

Du point de vue linguistique, la définition de l'analogie entre séquences reflète un biais indiscutable envers des situations dans lesquelles les données linguistiques sont distribuées selon un principe de co-variation d'un ensemble de préfixes et d'un ensemble de suffixes. Cette situation correspond effectivement à la distribution des formes lexicales selon des paradigmes flexionnels et dans une moindre mesure, dérivationnels, des langues indo-européennes. Dès lors que de nouvelles applications, par exemple à des tâches d'analyse syntaxique (voir A.0.8), ou de nouvelles familles de langues sont visées, il devient nécessaire de généraliser cette notion pour la rendre à même d'appréhender des données plus variées.

Un autre type d'extension s'impose : pouvoir considérer de nouvelles tâches demande de pouvoir manipuler des représentations (langages finis, arbres, etc) qui ne sont pas prises en compte dans le modèle d'inférence de la section 2.2, qui ne connaît, comme représentation structurée, que les séquences.

Ces différentes questions font l'objet des travaux présentés dans la section suivante.

2.4 Analogies généralisées

À la suite des travaux d'Yves Lepage [Lepage and Shin-Ichi, 1996, Lepage, 1998, 2000], j'ai proposé dans [Yvon, 2003, Stroppa and Yvon, 2004, Yvon et al., 2004] une généralisation de l'analogie entre séquences qui répond à certaines des limitations discutées dans la section précédente. Le point essentiel de cette généralisation est l'extension de la notion de proportion analogique à des configurations dans lesquelles les quatre termes de l'analogie se décomposent en un nombre quelconque de sous-chaînes. Cette première généralisation de l'analogie sur les mots fait l'objet de la section 2.4.1), dans laquelle j'évoque également la question du calcul de proportions avec des transducteurs finis.

Cet effort de reformulation et de formalisation des rapports de proportion entre séquences débouche presque naturellement sur deux généralisations plus fondamentales de cette notion : l'une qui l'étend à une large gamme de structures algébriques, en particulier à des structures traditionnellement utilisées en traitement automatique des langues (structures de traits, ensembles, langages, arbres) ; l'autre qui en dérive des notions graduelles de l'analogie entre mots, permettant de concevoir des mesures pour ordonner les solutions concurrentes d'une équation analogique (section 2.4.3).

2.4.1 Une nouvelle définition de l'analogie

Le point de départ de cette section est une nouvelle définition des proportions entre mots sur un alphabet fini, qui généralise celle de la section 2.2.1 pour prendre en compte l'alternance d'un nombre arbitraire de facteurs. Un rapport de proportion analogique les quatre mots x, y, z et t de Σ^+ est défini par :

Définition 2.3 (Proportion analogique)**.** $(x, y, z, t) \in \Sigma^+$ *constituent une proportion analogique, notée* $x : y :: z : t$ *si et seulement si* $\exists n > 0, \alpha_i, i = 1 \ldots n, \beta_i, i = 1 \ldots n, \in \Sigma^\star$ *tels que :*

$$\begin{array}{llll} \textit{soit} & x = \alpha_1 \ldots \alpha_n & t = \beta_1 \ldots \beta_n & y = \alpha_1\beta_2\alpha_3 \ldots \quad z = \beta_1\alpha_2\beta_3 \ldots \\ \textit{soit} & x = \alpha_1 \ldots \alpha_n & t = \beta_1 \ldots \beta_n & y = \beta_1\alpha_2\beta_3 \ldots \quad z = \alpha_1\beta_2\alpha_3 \ldots \end{array}$$

avec $\forall i, \alpha_i\beta_i \neq \epsilon$*. Le plus petit entier* n *pour lequel une telle relation existe est le* degré *de l'analogie.*

Un exemple de proportion vérifiant cette définition est :

subjectif : *subversif* :: *injection* : *inversion*

avec $n = 3$ et les facteurs suivants : α_1 =*sub*, α_2 =*ject*, α_3 =*if*, β_1 =*in*, β_2 =*vers*, β_3 =*ion*[21]. Le degré de cette proportion analogique est donc 3. D'une manière générale, le degré mesure la qualité intrinsèque d'une proportion : plus il est faible, meilleure semble l'analogie[22]. Comme la définition précédente, cette définition (i) est compatible avec les propriétés classiques de l'analogie données à la section 2.1 ; (ii) ne permet de garantir ni qu'une équation analogique possède toujours une solution, ni inversement l'unicité d'une solution[23].

2.4.2 Un solveur à états finis

[Yvon, 2003] montre comment ces différentes définitions de l'analogie entre mots peuvent s'implanter en utilisant le modèle des transducteurs finis. Les grandes lignes de cette construction sont également données dans [Yvon et al., 2004, Stroppa and Yvon, 2005a] ; elles sont esquissées dans les paragraphes ci-dessous, à partir des notions de mot complémentaire et de produit de mélange.

Mot complémentaire et mélange

Mots complémentaires Si u est un sous-mot de v, on appelle mot complémentaire de u par rapport à v le sous-mot formé en conservant uniquement les symboles de v qui ne sont pas dans u. Ainsi, par exemple *eeaie* est le complémentaire (vocalique) de *xmplr* par rapport à *exemplaire*. Formellement, si v est de longueur n, on note I_v l'ensemble ordonné d'indices $[1 \dots n]$, et I_u l'ensemble d'indices de I_v correspondant aux positions de v qui permettent de reconstruire u.

Définition 2.4 (Sous-mots complémentaires)**.** *Si u est un sous-mot de v, correspondant à l'ensemble d'indices I_u, l'ensemble des sous-mots complémentaires de u par rapport à v est défini par :*

$$v\backslash u = \{v(I_v \setminus I_u) \text{ avec } I_u \subset I_v \text{ et } v(I_u) = u\}$$

Si u n'est pas un sous-mot de v, $v\backslash u$ est vide.

À chaque mot v de $\Sigma^\star$ est associée une relation binaire (symétrique) sur $\Sigma^\star \times \Sigma^\star$, notée $\backslash_v$ et définie par :

Définition 2.5 (Complémentarité)**.** *$u\backslash_v w$ si et seulement si $u \in v\backslash w$.*

Pour tout mot v de $\Sigma^\star$, on note A_v l'automate canonique de v construit en établissant une bijection entre les états de A_v et les préfixes de v. Il est trivial d'en déduire un transducteur fini qui calcule la relation de complémentarité, en doublant dans A_v chaque transition par une transition ε et en ajoutant sur ces transitions les sorties adéquates (voir la figure 2.5).

Les notions de sous-mots complémentaires et de relation de complémentarité s'étendent directement à des langages quelconques ; il est immédiat de vérifier que si K et L sont rationels, l'ensemble des complémentaires de L par rapport au langage K est également rationnel.

[21] La décomposition privilégiée ici est une décomposition linguistiquement pertinente, mais il aurait été également possible de prendre α_2 =*jecti*, α_3 =*f*...

[22] Le cas limite est l'analogie de degré 0, qui met en relation quatre séquences identiques : $\forall u \in \Sigma^\star u : u :: u : u$.

[23] [Lepage, 2001a] donne une série de conditions nécessaires pour qu'une équation ait au moins une solution, conditions qui s'appliquent également ici. En particulier, si t est solution de $x : y :: z :?$, alors t contient tous les symboles de y et de z qui ne sont pas dans x, dans un ordre inchangé. Un corollaire est que toutes les solutions d'une équation analogique ont la même longueur.

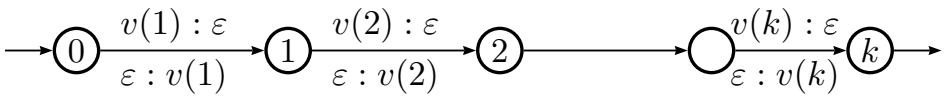

FIG. 2.5 – Transducteur fini calculant le complémentaire du langage $L = \{v\}$

Mélange Le *produit de mélange* ou mélange est défini, par exemple, dans [Sakarovitch, 2003] comme suit. Si u et v sont deux mots de $\Sigma^\star$, leur mélange $u \bullet v$ est le langage défini par :

$$u \bullet v = \{u_1 v_1 u_2 v_2 \dots u_n v_n, \text{ avec } u_i, v_i \in \Sigma^\star, u = u_1 \dots u_n, w = v_1 \dots v_n\}$$

Le mélange de u et v contient tous les mots formés des symboles de u et de v, avec la contrainte que si a précède b dans u ou v, alors cet ordre est respecté dans $u \bullet v$. Ainsi, par exemple, si l'on prend $u = abc$ et $v = def$, alors les mots suivants : $abcdef$, $abdefc$, $adbecf$ ... sont dans $u \bullet v$; ce n'est pas le cas de $abefcd$, dans lequel d suit e, alors qu'il devrait le précéder.

Cette opération s'étend aux langages de la manière suivante. Soient K et L deux langages :

$$K \bullet L = \bigcup_{x \in K, y \in L} x \bullet y$$

Le mélange est une opération rationnelle ; en particulier le mélange de deux mots est calculé par un automate résultant d'une forme particulière de produit des automates représentant ces mots [Sakarovitch, 2003].

Les notions de sous-mots et de mélange sont reliées par la relation suivante :

$$w \in u \bullet v \Leftrightarrow u \in w \backslash v$$

qui énonce que u et v sont en relation complémentaire par rapport à w si et seulement si w appartient au mélange de ces deux mots.

Un solveur analogique

Ces notions préliminaires étant posées, il devient possible de ré-exprimer la notion de proportion analogique. Le résultat principal est énoncé par la proposition suivante, démontrée dans [Yvon, 2003] :

Proposition 2.1.

$$x : y :: z : t \Leftrightarrow x \bullet t \cap y \bullet z \neq \emptyset$$

L'intuition de cette proposition est que, pour que l'analogie soit établie, il faut non seulement que les symboles de x et t soient les mêmes que ceux de y et z, mais également que les symboles de x (et de t) qui apparaissent dans y (et dans z) conservent leur ordonnancement. Le corollaire suivant s'en déduit immédiatement :

Proposition 2.2.

$$t \textit{ est une solution de } x : y :: z :? \Leftrightarrow t \in (y \bullet z) \backslash x$$

Ce résultat énonce que l'ensemble des solutions d'une équation analogique $x : y :: z :?$ est un ensemble rationnel, qui peut être calculé par un transducteur fini T. T est obtenu en construisant tout d'abord l'automate qui reconnaît le mélange de y et z, duquel on dérive ensuite l'automate des complémentaires, par une méthode analogue à celle exposée ci-dessus.

Deux résultats complémentaires sont établis dans [Yvon, 2003] :

– le solveur analogique implanté sous la forme d'un transducteur fini généralise strictement l'approche fondée sur des distances d'édition proposée par [Lepage, 1998] ;
– le calcul du degré d'une analogie $x : y :: z : t$ est réalisé en comptant dans t le nombre de transitions entre fragments de y et de z . Pour réaliser directement ce calcul, il suffit de composer T avec un transducteur fini qui effectue un tel décompte.

Ces divers résultats permettent d'implanter efficacement le calcul de *l'extension analogique* d'un langage fini L, notée L^3. L^3 contient tous les mots entrant dans une relation de proportion analogique avec 3 mots de L. Dans la pratique, plutôt que calculer explicitement le transducteur $(L \bullet L)\backslash L$, qui contient un nombre d'états quadratique par rapport à l'automate canonique de L, il est plus efficace de le simuler, à partir d'une représentation de A_L. Une étude plus complète de l'implantation de solveurs analogiques exploitant ce formalisme, qui inclut en particulier diverses optimisations permettant de trouver la proportion analogique de degré minimum entre quatre termes, est présentée dans [Stroppa, 2005, Stroppa and Yvon, 2006].

2.4.3 Compléments et généralisations

Le travail sur les analogies entre séquences présenté à la section précédente a donné lieu à deux types de développements, l'un susceptible de déboucher sur une version graduelle de l'analogie [Yvon et al., 2004] ; l'autre qui aboutit à une définition formelle de l'analogie valant pour une vaste gamme de structures algébriques couramment rencontrées dans les représentations linguistiques [Stroppa and Yvon, 2004, 2006]. Ces deux développements sont esquissés dans les paragraphes qui suivent, au terme desquels je conclurai sur les travaux actuellement en cours et sur quelques unes des perspectives qu'ils ouvrent.

Enchâssement d'analogies

Dans cette section sont introduites quelques notations et définitions supplémentaires, à partir desquelles est posée une troisième (et dernière) définition, encore plus générale que les précédentes, de l'analogie sur les mots, susceptible de fonder une version graduelle des proportions analogiques.

Alignements et analogies Soit '$-$' un nouveau symbole n'appartenant pas à Σ, $\overline{\Sigma}$ désigne $\Sigma \cup \{-\}$ et $\overline{\phi}$ la fonction qui associe les mots de $\overline{\Sigma}^*$ aux mots de $\Sigma^\star$ en supprimant les occurrences de '$-$'. Si x est un mot de $\Sigma^\star$, on note $\overline{\phi}^{-1}(x)$ l'ensemble des mots w de $\overline{\Sigma}^*$ tels que $\overline{\phi}(w) = x$.

Définition 2.6 (Alignement (binaire))**.** *Un* alignement *(binaire) entre x et y dans $\Sigma^\star$ est un mot u de $(\overline{\Sigma}\times\overline{\Sigma})^*$ tel que $\overline{\phi}(P_1(u)) = x$ et $\overline{\phi}(P_2(u)) = y$, où P_i désigne la projection canonique sur la ième composante de u.*

Ainsi, par exemple, $u = (c, g)(a, a)(-, m)(r, e)$ est un alignement entre les mots car et $game$, puisque $\overline{\phi}(P_1(u)) = \overline{\phi}(ca-r) = car$ et $\overline{\phi}(P_2(u)) = \overline{\phi}(game) = game$. Cette notion se généralise aux alignements d'arité quelconque.

Définition 2.7 (Alignement (n-aire))**.** *Un* alignement *n-aire entre n mots $x_1 \ldots x_n$ de $\Sigma^\star$ est un mot u dans $(\overline{\Sigma}^n)^*$ tel que $\forall i, \overline{\phi}(P_i(u)) = x_i$.*

Supposons maintenant que $\overline{R}$ est une relation sur des quadruplets de $\overline{\Sigma}^4$, qui spécifie les analogies valides *entre éléments de* $\overline{\Sigma}$. Il est possible d'utiliser $\overline{R}$ pour définir un prédicat $\overline{R}^*$ sur l'ensemble des alignements quaternaires :

$$\overline{R}^*(u) \text{ ssi } \forall i \in I_u, \overline{R}(u(i))$$

Ces préliminaires permettent alors de donner une nouvelle définition pour les proportions analogiques dans $(\Sigma^\star)^4$ par :

Définition 2.8 (Proportions analogiques généralisées)**.** $(x, y, z, t) \in \Sigma^\star$ *définissent une proportion analogique, notée x*:*y*::*z*:*t, si et seulement si il existe un alignement* u *de* x*,* y*,*z *et* t *tel que* $\overline{R}^*(u)$.

Analogies d'analogies Dans [Yvon et al., 2004], nous montrons que la définition (2.8) est formellement équivalente à la définition (2.3) pour un choix particulier de la relation $\overline{R}$, noté $\overline{R_0}$ et défini par :

$$\forall a, b \in \overline{\Sigma} \cup \{\varepsilon\}, \overline{R_0}(a,\ b,\ a,\ b) \text{ et } \overline{R_0}(a,\ a,\ b,\ b)$$

En d'autres termes, les analogies entre séquences telles que définies en (2.3) induisent, entre quadruplets de symboles des mots mis en relation, des *analogies élémentaires* qui sont, d'une certaine manière, triviales, puisqu'elles sont uniquement de la forme (a, a, b, b) ou (a, b, a, b).

Nous montrons ensuite que, *pour tout choix de* $\overline{R}$, la vérification et la résolution d'équations analogiques peuvent être implantées par des transducteurs finis, dont la construction est une généralisation de la construction de la section 2.4.2. Ce résultat permet d'envisager de travailler, sans perte signicative d'efficacité, avec des versions beaucoup plus « libérales » de la relation d'analogie (quelques exemples sont donnés ci-dessous), voire avec des versions qui abandonneraient certaines propriétés classiques de l'analogie, telle que, par exemple, la propriété d'échange des moyennes (définie à la section 2.1).

Dans le cas du traitement automatique des langues, diverses applications de ce travail sont envisageables. Ces résultats s'appliquent directement à des situations où les alphabets sur lesquels les mots sont définis sont munis d'une structure, qui permet de spécifier $\overline{R}$.

Par exemple, chaque symbole alphabétique usuel peut être vu comme possédant deux propriétés : sa valeur et sa casse. S'en déduit une définition pour $\overline{R}$, qui, en plus des relations de $\overline{R_0}$, inclurait des analogies élémentaires de la forme $\overline{R}(a, A, b, B)$. Il en va à l'identique si l'on prend pour alphabet Σ l'ensemble des symboles phonémiques, structuré par les valeurs des traits distinctifs assignés aux phonèmes. Dans ce contexte, R inclut des analogies élémentaires de la forme $\overline{R}(/\text{d}/,/\text{t}/,/\text{g}/,/\text{k}/)$, mettant en relation deux paires de phonèmes ne se distinguant que pour le trait de voisement. Au niveau des mots phonémiques, on obtient alors des analogies telles que : /tikə/ : /digə/ :: /pato/ : /bado/ (phonétiquement, *tique* est à *digue* comme *pataud* est à *badaud*). Ce type de modélisation vaut enfin pour un alphabet d'étiquettes morpho-syntaxiques qui serait structuré par des oppositions de nombre, de personne, de temps, de mode...

D'une manière plus générale, ce modèle permet de définir des analogies « emboîtées », dans lesquelles les analogies sur les mots au niveau n sont définies à partir de la relation d'analogie au niveau $n-1$. Il est alors possible de donner un sens à des analogies sur des séquences de séquences, et, grâce aux extensions introduites dans la section suivante, à des séquences d'arbres, des séquences de structures de traits, etc. Le coût algorithmique correspondant à la mise en œuvre de ces extensions reste toutefois à analyser plus précisément, avant de pouvoir envisager des expérimentations concrètes.

L'adoption d'une définition plus générale des analogies entre mots, se fondant sur une définition arbitraire des analogies élémentaires entre symboles, offre donc des perspectives applicatives nouvelles. L'augmentation du nombre de quadruplets définissant une proportion analogique, s'il peut servir à réduire le silence d'un système d'apprentissage par analogie, renforce corollairement le besoin pour des mesures scalaires de la qualité d'une analogie, qui permettraient d'ordonner les quadruplets solutions d'une équation.

Proportions analogiques généralisées

Ce travail étend la définition de l'analogie entre mots à l'ensemble des structures couramment rencontrées dans les descriptions linguistiques, et envisage le développement d'algorithmes efficaces pour réaliser les calculs de validation et de résolution d'analogie pour ces structures. Ce travail s'intéresse en premier

chef aux structures d'arbres, qui sont fréquemment rencontrées dans la description syntaxique (c.f. la section A.0.8). Des résultats préliminaires de ce travail sont parus dans [Stroppa and Yvon, 2004, Stroppa, 2005]. Pour les arbres, des éléments complémentaires sur les algorithmes de résolution d'analogies se trouvent dans [Stroppa and Yvon, 2006].

Cette section débute par une reformulation de la définition des proportions analogiques sur les mots, à partir de laquelle diverses généralisations sont ensuite considérées.

Algèbres, analogies Définissons tout d'abord la notion de *factorisation* d'un mot :

Définition 2.9 (Factorisation)**.** *On appelle factorisation de u toute séquence $u_1 \dots u_n$, avec $\forall i, u_i \in \Sigma^\star$, telle que : $u_1 \dots u_n = u$.*

La définition 2.3 se récrit alors sous la forme suivante :

Définition 2.10 (Proportion analogique, 2)**.** *$(x, y, z, t) \in \Sigma^+$ constituent une proportion analogique, notée $x : y :: z : t$ si et seulement s'il existe des factorisations $x_1 \dots x_n$, $y_1 \dots y_n$, $z_1 \dots z_n$, $t_1 \dots t_n$ respectivement de x, y, z et t telles que $\forall i, (y_i, z_i) \in \{(x_i, t_i), (t_i, x_i)\}$. Le plus petit entier n pour lequel de telles factorisations existent est le* degré *de l'analogie.*

Cette définition présente l'intérêt de mettre en évidence que l'analogie sur les séquences implique deux notions : la *décomposition* de chaque élément sous la forme d'un produit (ici, une concaténation) de termes plus petits, termes qui respectent des *contraintes d'alternances*. Cette définition se transpose directement dans le cadre plus général où l'on dispose d'un ensemble quelconque U muni d'une loi de composition interne, notée $\oplus$, en :

Définition 2.11 (Proportion analogique généralisée)**.** *$(x, y, z, t) \in U$ constituent une proportion analogique, notée $x : y :: z : t$ si et seulement s'il existe des factorisations $x_1 \oplus \dots \oplus x_n = x$, $y_1 \oplus \dots \oplus y_n = y$, $z_1 \oplus \dots \oplus z_n = z$, $t_1 \oplus \dots \oplus t_n = t$ telles que $\forall i, (y_i, z_i) \in \{(x_i, t_i), (t_i, x_i)\}$. Le plus petit entier n pour lequel de telles factorisations existent est le* degré *de l'analogie.*

Cette définition s'applique à de nombreuses structures algébriques classiques, telles que des espaces vectoriels, des groupes, etc. Ceci inclut en particulier des structures communément utilisées en traitement automatique des langues : ensembles munis de l'opération d'union ensembliste, structures de traits munies de l'opération de généralisation, etc.

Dans le cas où la structure sous-jacente possède des propriétés supplémentaires (commutativité, associativité, existence d'un élément neutre, existence d'un inverse unique pour $\oplus$), cette définition se simplifie et permet de retrouver des définitions bien connues de l'analogie. Ainsi, [Stroppa and Yvon, 2004] montre que dans le cas où $(U, \oplus)$ possède une structure de groupe :

Théorème 2.1 (Proportion analogique (dans un groupe))**.** *Si U est un groupe, alors $\forall x, y, z, t \in U$, $x : y :: z : t$ si et seulement si $x \oplus t = y \oplus z$.*

On retrouve ici une intuition classique, qui correspond au rapport de proportion dans $(\mathbb{N}, \times)$ ou dans un espace vectoriel.

Analogies entre langages La généralisation de l'analogie à des langages (en particulier à des langages finis) ouvre la possibilité de reproduire analogiquement des ensembles d'ambiguïtés, voire, plus généralement, de mettre en place des outils de traitement capables de prendre en compte de la variabilité (du

non-déterminisme) dans les sorties. Cette capacité fait partiellement défaut aux systèmes d'apprentissage fondés sur la classification (voir sur ce point la discussion de la section 2.1.4).

Considérons, par exemple, le cadre de l'apprentissage de la prononciation, et supposons que l'ensemble des prononciations de *médecin* contient les deux variantes {/medəsɛ̃/, /mɛtsɛ̃/}, que *cousine* se réalise avec ou sans e-muet final : {/kuzin/, /kuzinə/} et que *cousin* se prononce sans ambiguïté : /kuzɛ̃/. Supposons de plus que, pour la chaîne orthographique inconnue *médecine*, la proportion orthographique : *cousin* : *cousine* :: *médecin* : *médecine* ait été identifiée. Disposant d'une notion de l'analogie entre langages, il devrait être possible d'inférer pour *médecine* les variantes de prononciation suivantes : {/medəsinə/, /mɛtsinə/, /medəsin/, /mɛtsin/}.

Une définition possible pour l'analogie entre langages de $\Sigma^\star$ généralise directement la définition donnée sur les mots selon :

Définition 2.12 (Proportion analogique (pour des langages))**.** *$(X, Y, Z, T) \subset \Sigma^\star$ constituent une proportion analogique, notée $X : Y :: Z : T$ si et seulement s'il existe des factorisations*[24] *$X_1 \oplus \ldots \oplus X_n = x$, $Y_1 \oplus \ldots \oplus Y_n = Y$, $Z_1 \oplus \ldots \oplus Z_n = z$, $T_1 \oplus \ldots \oplus T_n = T$ telles que $\forall i, (Y_i, Z_i) \in \{(X_i, T_i), (T_i, X_i)\}$. Le plus petit entier n pour lequel de telles factorisations existent est le* degré *de l'analogie.*

Cette définition est cohérente avec l'exemple donné en préambule, avec $X_1 = \{$mɛt,medəs$\}$, $X_2 = \{$ɛ $\}$, $T_1 = \{$kuz$\}$, $T_2 = \{$ɛ̃,inə$\}$.

Le cas des arbres Les arbres constituent une représentation d'une importance cruciale pour représenter les structures récursives utilisées en morphologie, en syntaxe et en sémantique. Il n'est dans ce cas pas directement possible d'utiliser la définition (2.11), dans la mesure où il n'existe pas de loi de composition interne sur l'ensemble des arbres étiquetés qui permettrait simplement de décomposer un arbre comme produit d'arbres plus petits. L'opération de substitution à gauche, utilisée par exemple dans les travaux de R. Bod (e.g. [1995]) n'est pas associative et n'est donc pas directement compatible avec la définition (2.11).

Une manière alternative de traiter les arbres consiste à les représenter sous la forme de structures de traits récursives, représentation pour laquelle il existe déjà d'une définition de l'analogie. Si l'on adopte ce modèle, chaque nœud n_i, étiqueté par $lab(n_i)$ est représenté par une structure contenant un trait lab pour l'étiquette et k traits $d_1 \ldots d_k$, qui prennent pour valeur les structures associées aux nœuds dominés par n_i (voir la figure 2.6).

Le défaut principal de cette représentation est qu'elle conduit à distinguer (par des traits qui marquent l'ordre d'apparition des fils) des relations de domination qui sont pourtant identiques, débouchant sur une définition bien trop restrictive de l'analogie. Ainsi, par exemple, les quatre syntagmes de la figure 2.7 ne sont pas reconnus comme entrant dans un rapport de proportionnalité par une définition de l'analogie fondée sur les structures de traits et l'unification.

[Stroppa, 2005, Stroppa and Yvon, 2005b, 2006] présentent diverses propositions pour définir l'analogie entre arbres de manière à modéliser de telles relations, la dernière de ces références incluant également un algorithme pour résoudre les analogies entre arbres. Faute de disposer d'une implantation d'un solveur analogique pour les arbres, les expériences de [Stroppa and Yvon, 2005b] utilisent une linéarisation préalable des arbres et l'utilisation d'un algorithme de résolution sur les mots.

[24] Par souci de cohérence, le terme de *factorisation* d'un langage X continue à être réservé à des configurations dans lesquelles X s'écrit exactement comme la concaténation de $X_1 \ldots X_n$, bien que ce terme soit défini de manière différente par, par exemple, dans [Sakarovitch, 2003, p. 295], qui demande simplement $X_1 \ldots X_n \subset X$, avec chacun des X_i maximal au sens de l'inclusion ensembliste pour cette propriété.

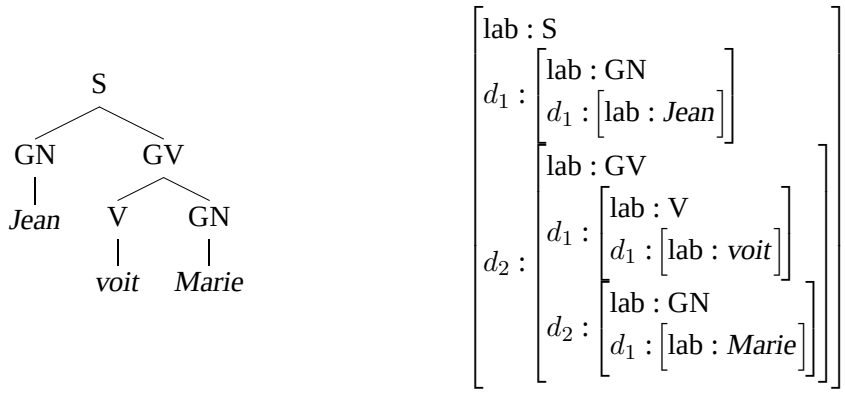

FIG. 2.6 – Représentation d'un arbre sous la forme d'une matrice de traits

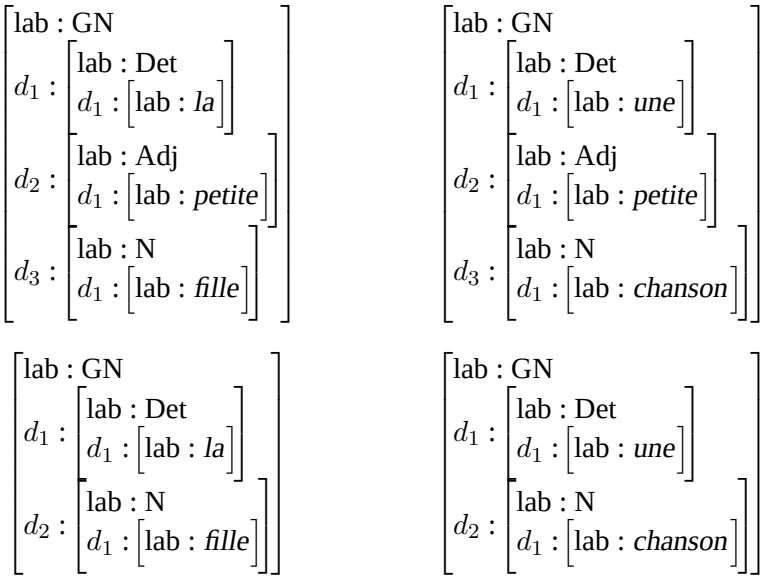

FIG. 2.7 – Une analogie « manquée »
'fille' (resp. 'chanson') apparaît sous le trait d_3 dans l'exemple du haut, sous le trait d_2 dans l'exemple du bas : le rapport analogique ne peut être capturé par unification entre ces structures.

2.5 Conclusion et perspectives

L'exposé des trois sections précédentes a permis de dresser un bilan de l'avancement de mes travaux théoriques et expérimentaux concernant l'apprentissage par analogie. J'ai ainsi successivement présenté le cadre général de l'apprentissage par analogie, s'appuyant sur une version simpliste de l'analogie entre séquences ; j'ai ensuite évoqué un certain nombre d'expérimentations autour des tâches d'acquisition de connaissances lexicales (prononciation et morphologie) ; puis proposé un cadre plus général pour l'analogie entre objets structurés, s'instanciant pour une large gamme de structures algébriques.

Dans cette section, j'évoque divers prolongements posssibles de ce travail : la poursuite des travaux expérimentaux, d'une part ; un approfondissement de la notion d'analogie graduelle, d'autre part.

2.5.1 Travail expérimental

Les avancées théoriques présentées à la section 2.4 ont permis de développer une boîte à outil à partir de laquelle de nombreuses expérimentations peuvent être entreprises. Dans le court terme, ce volet expérimental concernera en particulier les tâches suivantes :

– analyse morphologique automatique de mots inconnus, incluant notamment le calcul d'une représentation arborée, correspondant à un arbre d'analyse. Des résultats préliminaires utilisant une approximation de l'analogie entre arbres sont donnés dans [Stroppa and Yvon, 2005a,b].
  Concernant toujours l'apprentissage automatique, un enjeu théorique important consiste à tester son applicabilité à d'autres familles de langues que les langues indo-européennes, pour lesquelles les processus morphologiques sont décrits comme utilisant d'autres procédés que la simple concaténation d'affixes, en particulier les langues sémitiques.
– prononciation automatique, incluant : (i) des représentations phonémiques à base de structures de traits, généralisant l'approche présentée à la fin de la section 2.3.1 et permettant de faire des analogies « à deux niveaux » : le premier niveau correspond aux analogies entre structures de traits, et le second niveau aux analogies entre séquences de structures ; (ii) des variantes de prononciation, l'objectif étant de vérifier l'intérêt de pouvoir effectuer des analogies sur des langages finis pour apprendre à produire automatiquement de telles variantes ;
– d'étiquetage morpho-syntaxique : l'objectif étant de tester l'utilité des analogies à plusieurs niveaux, ici en représentant les traits morpho-syntaxiques sous la forme de structures multi-dimensionnelles.
– d'analyse syntaxique : l'avancement des réflexions sur les analogies entre arbres a conduit à proposer plusieurs définitions possibles des rapports de proportion, des études expérimentales doivent maintenant permettre de mieux évaluer les bénéfices et les limites de chacun de ces modèles.

2.5.2 Vers des modèles graduels de l'analogie ?

Une seconde perspective de recherche importante concerne l'intégration, dans un modèle qui reste essentiellement un modèle symbolique, d'un composant numérique, faisant le lien avec les modèles d'apprentissage statistiques/numériques. Il me semble en effet essentiel de pouvoir pondérer quantitativement les analogies extraites d'un lexique ou d'un corpus. Le bénéfice pratique d'une telle pondération est double :

– il permettrait de fournir la base à des heuristiques pour accélérer l'étape de recherche des triplets analogiques, par exemple en privilégiant l'exploration de zones du lexique qui sont analogiquement denses, ou encore la construction d'analogies impliquant des termes fréquents ;
– il permettrait également de fournir un modèle pour ordonner les sorties du modèle d'inférence, voire de définir une distribution de probabilités sur les sorties. Comme expliqué précédemment, une des faiblesses du système actuel reste son incapacité à sélectionner la bonne hypothèse, lorsqu'elle fait partie des hypothèses produites par le système ;

Comme discuté dans [Yvon et al., 2004], la définition (2.8) se prête à diverses extensions dans lesquelles les analogies sont dotées de valeurs scalaires mesurant la qualité d'une analogie.

Trois facteurs me semblent susceptibles d'intervenir dans la construction de telles mesures, qui sont clairement distingués dans le modèle présenté ci-dessus :

– un facteur lié à la « prototypicité » des termes entrant dans l'analogie : une analogie est d'autant plus satisfaisante qu'elle met en relation quatre objets qui sont eux-mêmes typiques. Une telle mesure pourrait, par exemple, dans le cas d'analogies dans le domaine lexical, s'appuyer sur la fréquence individuelle des lexèmes ou des formes participant à une proportion analogique.
– un facteur lié à la « décomposabilité » des termes entrant dans l'analogie : les définitions précédentes se fondent sur la modélisation des mots en tant qu'éléments d'un monoïde libre. Ainsi, les mots sont librement décomposables en la concaténation de symboles élémentaires d'un alphabet fini. La réalité linguistique est un peu plus complexe : comme discuté à la section A.0.6, la segmentation de séquences

linguistiques (des mots en graphèmes ou en morphèmes, des phrases en mots...) n'est pas toujours aussi triviale.

Une façon d'intégrer cette complexité consiste à associer aux décompositions d'un mot induites par leur implication dans un alignement quaternaire une valeur scalaire. Rappelons que, dans un alignement quaternaire, chaque composante de l'alignement est un mot de $\overline{\Sigma^\star}$. Pour chaque mot w mémorisé lors de l'apprentissage, il est possible d'associer aux éléments de $\overline{\phi}^{-1}(w)$ une mesure de la qualité de la segmentation induite. Pour en donner l'intuition sur un cas de tâche lexicale, pour un mot comme *technique*, ceci consisterait à noter la segmentation *techn-ique* comme supérieure à *te-ch-nique*, et incomparablement meilleure que *tec-hnique*.

– un troisième facteur intégrant les mesures de qualité des analogies élémentaires. L'apprenti analogique présenté dans les sections précédentes repose presqu'uniquement sur la définition des proportions entre mots, soit en dernière analyse, sur les analogies élémentaires (exprimées via la relation R_0) $a : a :: b : b$ et $a : b :: a : b$. Une piste naturelle pour étendre ce modèle consiste à apprendre ces analogies élémentaires, qui encodent la connaissance sur les alternances possibles ; une possibilité étant de chercher à compléter R_0 de façon à rendre les données d'apprentissage les plus redondantes possibles - en d'autres termes en recherchant les alternances de symboles élémentaires fréquemment représentées dans le lexique.

[Delhay and Miclet, 2004] s'inscrit dans une démarche similaire, en considérant des situations dans lesquelles l'alphabet Σ est nanti d'une distance entre symboles. Étant donnée cette distance, ces auteurs présentent une stratégie de résolution heuristique de la recherche d'une analogie de coût minimun.

Il reste, sur ces différentes questions, un important travail à la fois théorique, algorithmique et expérimental à conduire, en particulier pour compléter la définition des différents facteurs présentés ci-dessus, pour définir leur intégration dans un modèle cohérent et pour proposer des stratégies heuristiques permettant de calculer efficacement la ou les meilleures solutions d'une équation analogique.

Bibliographie

Agnar Aamodt and Enric Plaza. Cased-based reasoning : Foundational issues, methodological variants and system approaches. *Journal of Artificial Intelligence Communications*, 7(1) :39–59, 1994.

Steven Abney. Bootstrapping. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 360–367, Philadelphia, PS, 2002.

Stephen R. Anderson. *A-Morphous Morphology*. Cambridge University Press, 1992.

Mark Aronoff. *Morphology by itself*. The MIT Press, Cambridge, MA, 1994.

Paul Bagshaw. Phonemic transcription by analogy in text to speech synthesis : Novel word pronunciation and lexicon compression. *Computer Speech and Language*, 12 :119–142, 1998.

James P. Blevins. Word-based morphology. *Journal of Linguistics*, 42 :531–573, 2006.

Laurent Blin and Mike Edgington. Prosody prediction using a tree-structure similarity metric. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, volume III, pages 183–186, Beijing, China, 2000.

Laurent Blin and Laurent Miclet. Generating synthetic speech prosody with lazy learning in tree structures. In *Proceedings of CoNLL'00*, pages 87–90, Lisboa, PT, 2000.

Leopold Bloomfield. *Language*. Henry Bolt, New York, NY, 1933.

Rens Bod. *Enriching Linguistics with Statistics : Performance Models of Natural Language*. Institute for Logic, Language and Computation, Amsterdam, 1995.

Geert E. Booij. Paradigmatic morphology. In Bernard Tranel and Georgette Dal, editors, *Hommage à Danielle Corbin*, Amsterdam /Philadelphia, (to appear). John Benjamins.

Gavin Burnage. CELEX : a guide for users. Technical report, University of Nijmegen, Center for Lexical Information, Nijmegen, 1990.

Jaime G. Carbonell, Ryszard S. Michalski, and Tom M. Mitchell. An overview of machine learning. In Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell, editors, *Machine Learning : An Artificial Intelligence Approach*, volume I, pages 3–23, Palo Alto, CA, 1983. Tioga Publishing Company.

Claire Cardie. Empirical methods in information extraction. *AI magazine*, 18(4) :65–80, 1997.

David Chalmers, Robert French, and Douglas Hofstadter. High-level perception, representation, and analogy : A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, 4 :185–211, 1992.

Noam Chomsky. *Reflections on Language*. Pantheon House, New York, (trad. française : réflexions sur le langage, eds. flammarion, paris, 1981) edition, 1975.

Cecil H. Coker, Kenneth W. Church, and Mark Y. Liberman. Morphology and rhyming : two powerful alternatives to letter-to-sound rules. In *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France, 1990.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. MBT : A memory-based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 14–27, Copenhagen, Denmark, 1996.

Walter Daelemans, Antal Van Den Bosch, and Jakub Zavrel. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34 :11, 1999.

Georgette Dal. Analogie et lexique construit : quelles preuves ? *Cahiers de Grammaire*, 28 :9–30, 2003.

Robert I. Damper and Yannick Marchand. Improving pronunciation by analogy for text-to-speech applications. In *Proceeding of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

Robert I. Damper, Yannik Marchand, M.J. Adamson, and K. Gustafson. Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis. In *Proceeding of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

Mehdi Dastani, Bipin Indurkhya, and Remko Scha. An algebraic method for solving proportional analogy problems. In *Proceedings of MindII - Computational Models of Creative Cognition*, Dublin, Ireland, 1997. Dublin City University, School of Computer Applications.

Todd R. Davis and Stuart J. Russel. A logical approach to reasoning by analogy. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, pages 264–270, Milan, Italy, 1987.

Ferdinand de Saussure. *Cours de Linguistique Générale*. Payot, Paris, 2nde édition, 1967 edition, 1916.

Michael J. Dedina and Howard C. Nusbaum. PRONOUNCE : a program for pronunciation by analogy. *Computer Speech and Language*, 5 :55–64, 1991.

Élisabeth Delais. *Pour une approche parallèle de la structure prosodique : étude rythmique de l'organisation prosodique et rythmique de la phrase française*. PhD thesis, Université de Toulouse-le-Mirail, 1995.

Arnaud Delhay and Laurent Miclet. Solving analogical equations in sequences ; definition and resolution. In Michel Liquière and Marc Sebban, editors, *Proceedings of the International Conference on Grammatical Inference (ICGI 2004)*, pages 347–362, Montpellier, France, 2004. PUG.

Thomas G. Evans. A program for the solution of a class of geometric analogy intelligence test questions. In *Semantic information processing*, pages 271–353, Cambdridge, MA, 1968. The MIT Press.

Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. Structure mapping engine : algorithm and example. *Artificial Intelligence*, 41 :1–63, 1990.

Stefano Federici, Vito Pirrelli, and François Yvon. A dynamic approach to paradigm-driven analogy. In *Symbolic, connectionist, and statistical approaches to learning for natural language processing*, Berlin, 1996. Springer-Verlag.

Kenneth D. Forbus, Dedre Gentner, and Keith Law. MAC/FAC : A model of similarity-based retrieval. *Cognitive Science*, 19 :141–205, 1994.

Bernard Fradin. *Nouvelles approches en morphologie*. Presses Universitaires de France, Paris, France, 2003.

Robert M. French. The computational modeling of analogy making. *Trends in Cognitive Sciences*, 6(5) : 202–205, 2002.

Dedre Gentner. The mechanisms of analogical learning. In Vosniadou S. and Ortony A., editors, *Similarity and Analogical Reasoning*, pages 199–241. Cambridge University Press, 1989.

Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov, editors. *The analogical mind*. The MIT Press, Cambridge, MA, 2001.

Andrew R. Golding. *Pronouncing names by a combination of rule-based and case-based reasoning*. PhD thesis, Stanford University, Stanford, CA, 1991.

Nabil Hathout. Analogies morpho-synonymiques. une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. In *Proceedings of TALN'01*, volume 1, pages 223–232, 2001.

Jean-Paul Haton, Nadjet Bouzid, FrançoisCharpillet, Marie-Christine Haton, Brigitte Lâasri, Hassan Lâasri, Pierre Marquis, Thierry Mondot, and Amedeo Napoli. *Le raisonnement en Intelligence Artificielle*. Inter-éditions, 1991.

Robert R. Hoffman. Monster analogies. *AI Magazine*, pages 11–35, 1995.

Douglas Hofstadter. Epilogue : analogy as the core of cognition. In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov, editors, *The analogical mind*, pages 499–538, Cambridge, MA, 2001. The MIT Press.

Douglas Hofstadter and Melanie Mitchell. The copycat project : A model of mental fluidity and analogy-making. In Douglas Hofstadter and the Fluid Analogies Research group, editors, *Fluid Concepts and Creative Analogies*, pages 205–267. Basic Books, 1995.

Keith J. Holoyak and John E. Hummel. Understanding analogy within a biological symbol systems. In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov, editors, *The analogical mind*, pages 161–195, Cambridge, MA, 2001. The MIT Press.

Keith J. Holoyak and Paul Thagard. *Mental Leaps : Analogy in creative thought*. The MIT Press, Cambridge, MA, 1995.

Nancy Ide and Jean Veronis. MULTEXT : Multilingual text tools and corpora. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, volume 1, pages 588–592, Kyoto, Japan, 1994.

Bipin Indurkhya. Modes of analogy. In *Analogical and Inductive Inference*, number 397 in Lectures Notes in Artificial Intelligence, pages 217–229. Springer Verlag, 1989.

Esa Itkonen and Jussi Haukioja. A rehabilitation of analogy in syntax (and elsewhere). In András Kertész, editor, *Metalinguistik im Wandel*, pages 131–177. Peter Lang, 1997.

Daniel Jones. *Analogical Natural Language Processing*. UCL Press, London, 1996.

Mark T. Keane, Tim Ledgeway, and Stuart Duff. Constraints on analogical mapping : a comparison of three models. *Cognitive science*, 18 :387–438, 1994.

Boicho N. Kokinov. Integrating memory and reasoning in analogy-making : the AMBR approach. In Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov, editors, *The analogical mind*, pages 59–124, Cambridge, MA, 2001. The MIT Press.

Janet Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, CA, 1993.

Anne Lacheret-Dujour and Frédéric Beaugendre. *La prosodie du français*. Éditions du CNRS, Paris, France, 1999.

Joel Lachter and Thomas Bever. The relation between linguistic structure and associative theories of language learning : a constructive critic of some connectionist learning models. *Cognition*, 28 :195–247, 1988.

René Joseph Lavie. Le locuteur analogique ou la grammaire remise à sa place. Thése de doctorat de l'Université de Paris X Nanterre, 2003.

Yves Lepage. Languages of analogical strings. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, pages 488–494, Saarbrücken, Germany, 2000.

Yves Lepage. Défense et illustration de l'analogie. In *Actes de la 8eme conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, pages 373–377, Tours, France, 2001a.

Yves Lepage. Analogy and formal language. *Electronic notes in Theoretical Computer Science*, 47 :1–12, 2001b.

Yves Lepage. De l'analogie rendant compte de la commutation en linguistique. Mémoire d'habilitation à diriger des recherches. Université de Grenoble, 2003.

Yves Lepage. Lower and higher estimates of the number of true analogies contained in a large multilingual corpus. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, Genève, Suisse, 2004.

Yves Lepage. Solving analogies on words : An algorithm. In *Proceedings of COLING-ACL'98*, volume 2, pages 728–735, Montréal, Canada, 1998.

Yves Lepage. Open set experiments with direct analysis by analogy. In *Proceedings of NLPRS '99*, volume 2, pages 363–368, Beijing, China, 1999a.

Yves Lepage. Analogy+tables=conjugation. In G. Friedl and H.G. Mayr, editors, *Proceedings of NLDB'99*, pages 197–201, Klagenfurt, Germany, 1999b.

Yves Lepage and Ando Shin-Ichi. Saussurian analogy : A theoretical account and its application. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, pages 717–722, Copenhagen, Denmark, 1996.

Charles X. Ling and Marin Marinov. Answering the connectionist challenge : a symbolic model of learning the past tense of english. *Cognition*, 49(3) :235–290, 1993.

Brian MacWhinney and Jared Leinbach. Implementations are not conceptualizations : revising the verb learning model. *Cognition*, 40 :121–158, 1991.

Gary F. Marcus. The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, 56 :271–279, 1995.

Peter Hugoe Matthews. *Morphology*. Cambridge University Press, Cambridge, 1974.

Peter Hugoe Matthews. *Syntax*. Cambridge University Press, Cambridge, 1976.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

Sylvain Neuvel and Sean A. Fulop. Unsupervised learning of morphology without morphemes. In *Proceeding of the ACL Workshop on Morphological and Phonological Learning*, Philadelphia, PA, 2002.

Stephen Pinker and Alan Prince. On language and connectionism : analysis of a parallel distribution processing model of language acquisition. *Cognition*, 28 :73–194, 1988.

Vito Pirrelli and François Yvon. Analogy in the lexicon : a probe into analogy-based machine learning of language. In *Proceedings of the 6th International Symposium on Human Communication*, Santiago de Cuba, Cuba, 1999a.

Vito Pirrelli and François Yvon. The hidden dimension : paradigmatic approaches to data-driven natural language processing. *Journal of Experimental and Theoretical Artificial Intelligence (JETAI), Special Issue on Memory-Based Language Processing*, 11 :391–408, 1999b.

Tony A. Plate. Analogy retrieval and processing with distributed vector representations. *Expert systems*, 17 (1) :29–40, 2000.

Kim Plunkett and Virginia A. Marchman. Learning from a connectionist model of the acquisition of English past tense. *Cognition*, 61 :299–308, 1996.

David E. Rumelhart and James L. McClelland. On learning the past tenses of english verbs. In James L. McClelland and David E. Rumelhart, editors, *Parallel distributed processes : Explorations in the microstructure of cognition. Volume 2 : Psychological and biological models.*, pages 217–271. Bradford books/MIT Press, Cambridge, Massachussets, 1986.

Jacques Sakarovitch. *Eléments de théorie des automates*. Vuibert, Paris, 2003.

Ute Schmid, Helmar Gust, Kai-Uwe Kühnberger, and Jochen Burghardt. An algebraic framework for solving proportional and predictive analogies. In *Proceedings of the European Conference on Cognitive Science*, pages 295–300, 2003.

Patrick Schone and Dan Jurafsky. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, Lisbon, Portugal, 2000.

Patrick Schone and Dan Jurafsky. Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, Pittsburgh, PA, 2001.

Royal Skousen. *Analogical Modeling of Language*. Kluwer, Dordrecht, 1989.

Royal Skousen, Deryle Lonsdale, and Dilworth B. Parkinson, editors. *Analogical Modeling*. John Benjamins Publishing company, Amsterdam/Philadelphia, 2002.

Andrew Spencer. *Morphological Theory*. Blackwell, Oxford, UK, 1991.

Nicolas Stroppa. *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. PhD thesis, École Nationale Supérieure des Tél'ecommunications, Paris, 2005.

Nicolas Stroppa and François Yvon. Analogies dans les séquences : un solveur à états finis. In *Actes de TALN 2004*, Fez, Marocco, 2004.

Nicolas Stroppa and François Yvon. Analogical learning and formal proportions : Definitions and methodological issues. Technical Report ENST-D004, École Nationale Supérieure des Télécommunications, Paris, France, 2005a.

Nicolas Stroppa and François Yvon. Formal models of analogical relationships. Technical report, à paraître, ENST, Paris, France, 2006.

Nicolas Stroppa and François Yvon. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pages 120–127, Ann Arbor, MI, 2005b.

Paul Thagard, Keith J. Holoyak, Greg Nelson, and David Gochfeld. Analog retrieval by constraint satisfaction. *Artificial Intelligence*, 46(3) :259–310, 1990.

Peter D. Turney. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3) : 251–278, 2005.

Peter D. Turney. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL-06)*, pages 313–320, Sidney, Australia, 2006.

Peter D. Turney and Michael L. Littman. Learning analogies and semantic relations. Technical Report NRC-46488, National Research Concil Canada, 2003.

Antal van den Bosch. *Learning to pronounce written words : A study in inductive language learning*. PhD thesis, University of Maastricht, 1997.

Antal van den Bosch and Walter Daelemans. Memory-based morphological processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 285–292, College Park, MD, 1999.

Jorn Veenstra, Antal van den Bosch, Sabine Buchholz, Walter Daelemans, and Jakub Zavrel. Memory-based word sense disambiguation. *Computers and the Humanities, special issue on Senseval, Word Sense Disambiguation*, 34(1-2) :171–177, 2000.

Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1) :168–173, 1974. ISSN 0004-5411. doi : http://doi.acm.org/10.1145/321796.321811.

François Yvon. Finite-state transducers solving analogies on words. Technical Report D008, Ecole Nationale Supérieure des Télécommunications, Paris, France, 2003.

François Yvon. *Prononcer par analogie : motivations, formalisations et évaluations*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 1996.

François Yvon. Pronouncing unknown words using multi-dimensional analogies. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 199–202, Budapest, Hungary, 1999.

François Yvon, Nicolas Stroppa, Arnaud Delhay, and Laurent Miclet. Solving analogies on words. Technical Report D005, Ecole Nationale Supérieure des Télécommunications, Paris, France, 2004.

Jakub Zavrel and Walter Daelemans. Feature-rich memory-based classification for shallow nlp and information extraction. In Jurgen Franke, Gholamreza Nakhaeizadeh, and Ingrid Renz, editors, *Text Mining, Theoretical Aspects and Applications*, pages 33–54. Springer Physica-Verlag, 2003.

Jakub Zavrel, Walter Daelemans, and Jorn Veenstra. Resolving PP attachment ambiguities with memory-based learning. In Mark Ellison, editor, *Proceedings of the workshop on Computational Natural Language Learning (CoNLL'97)*, pages 136–144, Madrid, 1997.

Chapitre 3

Autour de l'exploitation statistique de corpus

Ce chapitre présente une synthèse de mes travaux sur l'utilisation de méthodes statistiques pour l'exploitation de corpus textuels. Mes contributions sur ce thème s'organisent en deux volets principaux.

D'une part des travaux portant sur l'apprentissage et l'exploitation de modèles de langage (une présentation de la tâche est donnée à la section A.0.8), qui trouvent un débouché en traitement automatique de la parole, notamment en reconnaissance vocale. Plusieurs thèses sont actuellement en cours sur ce thème :

– la thèse d'Antoine Ghaoui, co-encadrée avec G. Chollet (ENST Paris) et C. Mokbel (Université de Balamand au Liban), porte sur la construction de modèles statistiques de langage pour la reconnaissance automatique de l'arabe. Elle considère des variantes des modèles k-grammes qui soient plus adaptées aux spécificités (en particulier morphologiques) de la langue arabe. Des résultats préliminaires de ces travaux sont présentés dans [Ghaoui et al., 2005].
– la thèse d'Aleksandra Krul, qui se déroule dans le cadre d'un contrat avec France Télécom R&D, en collaboration avec T. Moudenc et G. Dammati. Cette thèse, débutée en 2005, porte sur l'utilisation des modèles de langage pour des applications de synthèse vocale par sélection d'unités. On trouvera dans [Krul et al., 2006] des résultats préliminaires concernant une tâche de sélection automatique d'un corpus optimal.
– la thèse de Lin Shiuan-Sung porte enfin sur la mise en œuvre de techniques d'apprentissage discriminant pour les modèles de langage, dans un contexte de reconnaissance vocale.

Les plus aboutis de ces travaux, en particulier ceux conduits en collaboration avec Lin Shiuan-Sung, seront évoqués dans le chapitre 4 consacré aux contributions au traitement automatique de la parole, plus précisément à la section 4.2.5.

L'autre volet de mes travaux porte sur les techniques de fouille de texte dans le contexte de la gestion documentaire : recherche et extraction d'information, catégorisation et classification de documents. Ils ont également donné lieu à la direction ou co-direction de trois thèses, en collaboration avec des partenaires industriels. Ce sont ces travaux que j'ai choisi de présenter ici, en revisitant successivement une application d'acquisition de ressources sémantiques (section 3.1), puis des contributions à la catégorisation (section 3.2 et à la classification de documents (section 3.3).

Dans la mesure où, pour l'essentiel, ces contributions reposent sur des modèles statistiques bien documentés, auxquels nous avons accordé une large place dans le chapitre 1, nous nous en tenons ici à une mise en contexte minimale. Concernant les deux derniers thèmes, on trouvera des éléments complémentaires dans les documents joints, notamment dans les références [Vinot et al., 2003] et [Rigouste et al., 2006a].

3.1 Acquisition de paraphrases

Ce travail[1] s'inscrit dans une longue lignée d'études concernant l'acquisition de ressources lexicales, notamment de relations de similarité sémantique entre mots à partir de l'analyse distributionnelle de corpus. Le principal mérite de ce travail très empirique est de s'attaquer à un « corpus » particulier, à savoir le Web, dont la redondance est exploitée à travers une boucle d'apprentissage *presqu'entièrement non-supervisée*. On trouvera une description plus complète de l'évolution de ce travail dans les diverses publications qui ont jalonné son avancement [Duclaye et al., 2002b, 2003a,b,c].

3.1.1 Cadre général

Le cadre applicatif de ce travail est l'amélioration d'une application de question-réponse, destinée à rechercher sur le Web la réponse à des questions factuelles. Les systèmes de question-réponse (QA) améliorent les systèmes classiques de recherche documentaire par deux aspects : ils sont capables (ii) d'analyser des demandes en langage naturel, dont la complexité excède largement celle des requêtes traditionnelles (mots-clés) et (ii) de retourner non pas des documents mais de courts fragments contenant (idéalement contenant uniquement) la réponse précise à la question posée. La mise en place d'une « piste » portant sur ces systèmes au sein de la conférence TREC (à partir de 1999) a conduit au développement de nombreux systèmes pour cette tâche (on se reportera aux actes de TREC[2] pour les détails sur cette tâche et pour se faire une idée des évolutions des systèmes de QA).

Le système qui a servi de toile de fond à cette étude possède une architecture relativement classique intégrant des modules :
(a) d'analyse de la question et de construction de patrons d'extraction des réponses ;
(b) de recherche sur le Web de documents pertinents ;
(c) d'extraction des réponses par application des patrons d'extraction ;
(d) de visualisation des réponses ;
Une description plus détaillée de l'architecture de ce système est donnée dans [Duclaye et al., 2002a].

Dans ce contexte, notre travail s'est focalisé sur la question de l'acquisition automatique de quasi-synonymes ou plus généralement de paraphrases[3] pour des concepts relationnels exprimés par des verbes ou des groupes verbaux. Ces paraphrases permettent d'améliorer l'étape de constitution des patrons d'extraction utilisés dans l'étape de recherche des réponses (étape (c)), ainsi que l'étape de recherche documentaire (b).

Pour illustrer cette idée, supposons que la question posée au système soit : *Qui a écrit Moby Dick ?* Les réponses à cette question sont obtenues en recherchant dans les documents pertinents des propositions dont la tête verbale est une forme de ÉCRIRE, ayant pour objet direct *Moby Dick*. Il existe bien d'autres manières d'exprimer la relation entre un auteur et son œuvre, qui utilisent :
– des procédés syntaxiques et morphologiques réguliers (passivation, nominalisation, etc.)
– d'autres verbes qui peuvent être soit des synonymes plus ou moins proches (*rédige, compose*, etc.), soit des implications valides dans ce contexte (*s'attelle, travaille à, se consacre à*, etc.)
– des procédés typo-dispositionnels (par exemple, dans une liste d'ouvrage *Moby Dick (Herman Melville, 1851)*).
Ce travail vise donc à acquérir de manière non-supervisée de telles reformulations d'une relation sémantique. Les reformulations ainsi produites étant destinées à être manuellement validées avant intégration au système de Question/Réponse sous la forme principalement de motifs d'extraction.

[1] Cette section est une synthèse du travail réalisé en collaboration avec le centre de recherche de France Télécom dans le cadre de l'encadrement du travail de thèse de Florence Duclaye, effectué conjointement avec Olivier Collin (FT R&D).
[2] Voir `http://trec.nist.gov/data/qa.html`.
[3] Entendu ici dans un sens relativement large, il s'agit en fait souvent plus de relations d'implication (*entailment*).

Il s'inscrit dans un axe de recherches autour de l'acquisition de relations sémantiques en corpus visant à la construction de ressources lexicales. Les travaux précurseurs en la matière sont certainement ceux de M. Hearst [1992], qui s'intéresse à un type de relation particulière, la relation d'hyponomie : dans son travail, les concepts impliqués dans cette relation sont extraits par analyse de quelques patrons textuels prédéfinis. Ainsi, de l'observation répétée de *X is a sort of Y*, *X is a Y such that*, on peut déduire que le concept référé par X est un sous-type de celui dénoté par Y. Cette approche a donné lieu à de multiples extensions, accompagnant le développement de besoins de ressources lexicales pour des multiples domaines (en particulier pour des réseaux sémantiques). [Turney, 2005] en constitue une extension particulièrement aboutie, dans la mesure où il considère un très nombre de patrons textuels (64), dont la fréquence en corpus permet implicitement de caractériser une paire de concepts, de la même manière qu'un vecteur de co-occurrences permet de caractériser implicitement le ou les sens d'un mot. Étant donnée une telle représentation vectorielle de paires de concepts, il devient possible de mettre en œuvre tout l'éventail des techniques classiques utilisées dans le modèle vectoriel en recherche d'information : calcul de similarité par le cosinus, analyse sémantique latente des relations [Turney, 2006], etc.

L'idée principale de notre approche est de chercher à tirer parti de la redondance des informations disponibles sur le Web, à l'instar de nombreux autres travaux, en particulier de Brin [1998] (voir également [Kilgariff and Grefenstette, 2003] et les articles publiés dans le même volume pour une présentation de la problématique général du Web comme corpus ou encore [Turney, 2005]). D'un point de vue technique, on peut y voir un précurseur de [Turney, 2006], faisant appel à PLSA [Hofmann and Puzicha, 1998], plutôt qu'à LSA [Deerwester et al., 1990].

3.1.2 Apprentissage de relations de paraphrase

Principe général

D'un point de vue méthodologique, notre démarche s'inspire essentiellement des techniques d'apprentissage semi-supervisé [Blum and Mitchell, 1998, Riloff, 1999], ainsi que des nombreux travaux visant à construire des classes de mots synonymes ou sémantiquement apparentés (en particulier [Rooth, 1995, Lin, 1998]). Comme ces derniers, elle se fonde sur l'hypothèse distributionnelle classique qui s'exprime ici par le fait que deux formulations sont des paraphrases potentielles si elles apparaissent fréquemment avec les mêmes arguments.

Cette hypothèse est exploitée dans une boucle d'apprentissage qui alterne des étapes de recherche sur le Web ((i) et (iii)) et des étapes de filtrage des paraphrases découvertes ((ii) et (iv)). Ces différentes étapes s'organisent globalement comme suit (on se reportera aux références citées ci-dessus pour une description complète) :

(i) disposant d'un ensemble de formulations d'une relation sémantique donnée, des requêtes sont construites et sont envoyées à un moteur de recherche. Les fragments pertinents, correspondant à des propositions complètes utilisant ces formulations, en sont extraits.

(ii) L'ensemble des couples (prédicats, arguments) obtenu en (i) est représenté dans une matrice de co-occurrences, qui sert de base à une procédure de classification non-supervisée des lignes (les prédicats) et des colonnes (les arguments). Cette procédure est explicitée ci-après.

(iii) une recherche est lancée à partir des nouvelles paires d'arguments extraites en (ii), permettant d'étendre le corpus des paraphrases potentielles.

(iv) comme en (ii), la nouvelle matrice de co-occurrences est analysée, permettant d'identifier de nouvelles paraphrases potentielles, déclanchant un nouveau cycle de recherche/filtrage (retour en (i)).

L'exécution de ce cycle d'acquisition est déclenché par *une unique formulation initiale*, qui est celle pour laquelle on recherche des paraphases. Cette formulation constitue la seule information de supervision fournie au système, le reste du processus se déroulant intégralement en mode non-supervisé.

Filtrage des paraphrases extraites

Les étapes de filtrage s'appuient sur un ensemble automatiquement construit de couples (prédicats, arguments), qui sont représentés dans une matrice de co-occurrence : chaque cellule de cette matrice enregistre le nombre d'occurrences d'une formulation avec un ensemble d'arguments.

Parmi les multiples outils disponibles pour l'analyse de cette matrice, nous avons adopté un modèle probabiliste de mélange de type PLSA [Hofmann and Puzicha, 1998], qui nous permet d'agréger incrémentalement de manière relativement simple les informations acquises au cours des étapes successives de construction du corpus.

Dans ce modèle, on suppose que prédicats et couples d'arguments prennent leurs valeurs dans deux ensembles finis $\mathcal{X}$ et $\mathcal{Y}$. La génération d'un couple (x, y), où x dénote la valeur du prédicat et y celle de son couple d'arguments s'opère en deux étapes :
– (i) tirage d'une variable sémantique latente S suivant une distribution discrète $\Pr(s), s \in \mathcal{S}$.
– (i) conditionnellement à la réalisation de S, tirage indépendant d'une valeur pour le prédicat et d'une valeur pour les arguments selon des distributions $P(x|s), x \in \mathcal{X}$ et $P(y|s), y \in \mathcal{Y}$.

Soit un corpus $\mathcal{C} = \{(x^{(i)}, y^{(i)}), i = 1 \dots n\}$ constitué de n paires i.i.d. Étant donné une valeur θ pour l'ensemble des paramètres, la log-vraisemblance du corpus s'écrit :

$$\mathcal{L}(\mathcal{C}|\theta) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{N}(x, y) \log(P_\theta(x, y)) \tag{3.1}$$

$$= \sum_{x} \sum_{y} \mathrm{N}(x, y) \log(\sum_{s} P(s)P(x|s)P(y|s)) \tag{3.2}$$

avec $\mathrm{N}(x, y)$ le nombre d'occurrences de la paire (x, y) dans le corpus. La maximisation directe de la quantité (3.2) n'étant pas possible, l'estimation des paramètres passe par la mise en œuvre d'un algorithme de type EM (Expectation-Maximisation). On se reportera par exemple à [Hofmann and Puzicha, 1998] pour une présentation plus complète de ce modèle, incluant en particulier les équations de remise à jour des paramètres.

Dans notre contexte, S est une variable binaire : (x, y) est ou n'est pas un couple représentant la relation sémantique cherchée. Au terme de la convergence de l'algorithme EM, un test statistique simple comparant les probabilités *a posteriori* $P(S = 1|(x, y))$ et $P(S = 0|(x, y))$ qu'un prédicat (ou des arguments) représentent ou non la relation sémantique visée permet de sélectionner (par seuillage) de nouveaux termes, qui font l'objet d'une nouvelle phase de recherche sur le Web.

L'évolution des paramètres du modèle au fur et à mesure que le corpus s'étend est gérée par l'artifice suivant : à la première étape de filtrage, l'initialisation des paramètres est uniforme (sauf pour l'unique exemple positif). Lors des itérations ultérieures, les paramètres calculés lors d'une étape de filtrage t peuvent servir à l'initialisation de l'étape $t + 1$, permettant d'obtenir une stabilité dans les reformulations qui sont progressivement extraites.

Quelques conclusions

Un prototype réalisant l'ensemble de ce programme a été implanté et testé dans le cadre du travail de thèse précité. Les principaux acquis des expériences menées à l'aide de cet outil sont les suivants.
– Dans la majorité des exemples testés, la procédure d'acquisition converge effectivement (c'est-à-dire que le système cesse, typiquement au bout d'une dizaine d'itérations, de proposer de nouvelles formulations potentiellement intéressantes). Ce comportement n'est toutefois pas garanti : en particulier, l'existence de reformulations très générales (par exemple *travaille à* pour la relation d'autorat) pouvant conduire le système à diverger. Diverses propositions sont faites dans [Duclaye, 2004] pour remédier à ce problème, en particulier pour mieux spécifier les requêtes en y intégrant des termes généraux du domaine étudié.

– Les reformulations extraites, du moins celles dont le score est élevé, sont le plus souvent correctes. À défaut de paraphrases, le système extrait également fréquemment des termes qui sont thématiquement proches de la cible.
– la procédure de construction de la matrice de co-occurrences se fonde sur une analyse de surface (*chunking*) des fragments retrouvés sur le Web pour identifier les couples arguments-prédicats. Cette analyse s'est avérée insuffisamment précise et de nature à renforcer le caractère extrêmement creux des matrices de co-occurrence extraites, ce qui nuit à la convergence globale de l'apprentissage.
– L'exploitation de données collectées sur le Web est une tâche particulièrement ingrate et frustrante : l'obligation de passer par les portails publics soumet le développement d'outils de traitements de corpus à de multiples vexations (changement des formats de présentation, instabilité des réponses à une requête, limitation du nombre de requêtes dans une fenêtre temporelle donnée), sans mentionner les difficultés inhérentes au traitement de données extraites du Web (multilinguisme, spams, etc.). Ces problèmes d'ingénierie rendent d'autant plus nécessaires les initiatives visant à construire de très grands corpus (qu'ils soient ou non issus du Web) en accès libre à destination d'une utilisation en traitement automatique des langues.

Au final, ce travail a permis de montrer la faisabilité de stratégies d'apprentissage faiblement supervisées, visant à exploiter la très grande redondance des informations disponibles sur le Web, ainsi qu'à mettre en évidence certaines des difficultés de cette approche (disponibilité d'outils robustes de pré-traitement et de traitement linguistiques, problèmes liés à l'utilisation du Web comme corpus, etc.). Pour conclure, signalons que ce travail a donné lieu au développpement d'un système opérationnel, pour lequel une version de production est en cours de développement.

3.2 Catégorisation automatique de documents

Cette section présente une étude[4] sur la catégorisation de documents, tâche emblématique et largement documentée de la fouille de textes. L'originalité principale de ce travail expérimental est de s'inscrire dans un cadre applicatif particulier, qui implique notamment l'intervention d'opérateurs humains. Ce contexte nous a permis d'une part de montrer que, pour autant que l'application s'y prête, des modèles génératifs très simples (ici un catégoriseur fondé sur Rocchio) peuvent s'avérer plus utiles que des apprentis plus sophistiqués (par exemple les SVMs) ; il nous a également donné l'occasion de nous pencher sur un problème peu étudié, à mi-chemin entre classification supervisée et non-supervisée et de montrer l'intérêt de techniques de clustering y compris dans un cadre de catégorisation.

Après une rapide présentation de la tâche et de l'état de l'art, nous décrivons plus en détail l'application sur laquelle nous avons travaillé, ainsi que les principaux résultats de notre étude.

La catégorisation de documents

Dans sa forme la plus simple, la catégorisation de documents (voir [Sebastiani, 2002, Vinot, 2004] pour des états de l'art récents) consiste à assigner à un document une ou plusieurs étiquettes catégorielles permettant d'indexer le document dans un ensemble prédéfini de thèmes, les thèmes pouvant être organisés selon une hiérarchie. Originellement conçue pour assister le classement documentaire d'ouvrages ou d'articles dans des domaines techniques ou scientifiques, la catégorisation de documents est devenue une tâche de base du traitement de documents[5] textuels ou sonores et se décline sous de nombreuses variantes. Notons principalement, en plus de l´indexation thématique « classique », les tâches de :

[4] J'ai été conduit à m'intéresser à cette tâche principalement à l'occasion de la supervision du travail de thèse de Romain Vinot. L'exposé qui suit lui doit naturellement beaucoup.

[5] À prendre ici dans un sens particulièrement large : selon les applications, un document peut être un livre, le contenu complet d'un site Web, un email, ou encore la transcription automatique d'un énoncé isolé dans un système de dialogue.

– *routage* : pour chaque document, il s'agit d'effectuer une classification multi-classe permettant de transmettre le document vers le ou les récepteur(s) le(s) plus approprié(s) ; ceci valant également pour l'adressage de requêtes écrites ou orales vers le bon destinataire ;
– *filtrage* : pour chaque document, il s'agit d'effectuer une décision binaire concernant la pertinence du document pour un destinataire particulier. Un cas d'école est le filtrage automatique de pourriels (ou spams) [Androutsopoulos et al., 2000], qui a donné lieu à de nombreuses implantations opérationnelles de systèmes de filtrage ;
– *suivi d'événements* : à partir d'un petit nombre de documents relatant un événement particulier, il s'agit de trouver, dans une large collection de textes, les documents traitant du même événement ;

Notons, pour achever ce bref panorama, qu'il existe de nombreux contextes applicatifs dans lesquels les catégories ne sont pas interprétables comme des thèmes : ainsi, par exemple, les tâches consistant à classer les documents par auteur, par genre, par style, par langue, ou encore selon que le document exprime un jugement positif ou négatif [Pang et al., 2002, Vinot, 2004, Pang and Lee, 2005].

Durant les années 90, la catégorisation de documents a été reformulée comme un problème d'apprentissage et est rapidement devenue un banc d'essai recommandé pour les techniques de catégorisation : classifieur bayésien naïf [McCallum and Nigam, 1998], Machines à vecteurs de support [Joachims, 1998], le boosting [Schapire et al., 1998, Schapire and Singer, 2000] et bien d'autres. En dépit de ses performances relativement médiocres dans toutes les évaluations publiées (voir eg. [Yang, 1997]), le classifieur bayésien naïf reste un des outils de catégorisation de documents les plus populaires et est aujourd'hui intégré[6] à de nombreux produits commerciaux.

Cette reformulation se fonde sur une représentation simplifiée des documents, qui modélise un document par un vecteur dont la valeur de chaque composante est une fonction du nombre d'occurrences d'un terme particulier. Le choix de l'ensemble des termes (ou multi-termes) appropriés, d'une méthode de calcul des composantes, etc., conditionnent largement les performances des différents systèmes de catégorisation, qui se répartissent en deux grandes familles :
– ceux qui modélisent directement les fréquences d'occurrences (modèles probabilistes) ;
– ceux qui exploitent une transformation des fréquences en valeurs numériques réelles, représentant chaque document comme un vecteur dans $\mathbb{R}^n$ (modèles vectoriels). Cette opération permet directement d'utiliser sur des documents les techniques traditionnelles de catégorisation automatique.

Quelle que soit l'approche retenue, un des traits particuliers de cette tâche est la très grande dimensionnalité de l'espace dans lequel les documents sont représentés, qui comprend typiquement plusieurs milliers ou dizaines de milliers de dimensions.

3.2.1 Cadre général

La problématique générale de ce travail visait initialement au développement d'un système de réponse semi-automatique à des questions ou requêtes adressées par courriel, s'intégrant dans un progiciel commercial de suivi de la correspondance client (dit *mail center*, par analogie avec le *phone center*).

Dans un premier temps, nous avons montré que ce problème pouvait être en partie reformulé et traité de manière efficace comme un problème de catégorisation de documents : supposant donnée une base de réponses typiques, validées par un opérateur, la construction automatique d'une réponse peut, en effet, être envisagée comme l'affectation d'un mail à la catégorie correspondant à la réponse typique la mieux appropriée pour le problème mentionné dans le courriel. Les cinq catégories les plus pertinentes sont alors proposées à un opérateur, qui peut alors décider de renvoyer une des réponses proposées, de l'adapter, ou encore d'ignorer les propositions du classifieur et de composer une nouvelle réponse. Des études empiriques ont montré qu'environ 50% des courriels reçus pouvaient recevoir une des réponses prédéfinies, sans autre intervention

[6] Sous une forme pas toujours très orthodoxe.

de l'opérateur.

Après examen empirique de plusieurs systèmes de catégorisation de textes [Yang, 1997, Lewis, 1998, Joachims, 1998], cette démarche a été implantée en utilisant un apprenti simple, généralement désigné sous le nom de Rocchio, en référence aux travaux originaux de J. Rocchio en recherche documentaire [1971]. Ce système de catégorisation se rattache à la famille des catégoriseurs vectoriels et construit des séparations entre classes de la manière suivante : chaque document d est représenté par un vecteur normé $\mathbf{x}_d$, dont chaque composante x_{dw} est fonction de la fréquence du terme w dans le document d. Dans notre système, le calcul de ces composantes n'implique aucun pré-traitement : toutes les formes recensées dans le corpus sont conservées et sont associées à une composante de $\mathbf{x}_d$. Chaque catégorie $c \in \mathcal{C}$ est représentée par un vecteur *prototypique* $\mathbf{c}$ calculé selon :

$$\mathbf{c} = \frac{t}{n_c} \sum_{d \in c} \mathbf{x}_d - \frac{1-t}{n_d - n_c} \sum_{d \notin c} \mathbf{x}_d \tag{3.3}$$

avec n_d le nombre total de documents, n_c le nombre de documents dans la classe c et $t \in [0 \dots 1]$ un paramètre libre du modèle. Ces vecteurs prototypiques sont ensuite normés. La catégorie d'un nouveau document est simplement déterminée en calculant sa similarité (son cosinus) avec les différents prototypes existants, la classe choisie étant celle du prototype le plus proche.

Pour le cadre applicatif du *mail center*, nous avons montré [Vinot and Yvon, 2001, 2002] que cet apprenti réunit de nombreux avantages :

– la capacité à gérer directement des problèmes multi-classe avec un nombre de classes potentiellement élevé ;
– la possibilité de mettre à jour incrémentalement[7] les paramètres (les positions des prototypes) en réaction aux actions des opérateurs ou de créer de nouvelles catégories, tout en maintenant le système relativement stable. La stabilité du classifieur obtenu est, en effet, une propriété importante pour l'acceptabilité du système par des opérateurs ;
– la rapidité du calcul de la catégorisation, permettant de gérer efficacement de très gros flux de données ;
– la capacité à construire une règle de décision acceptable à partir d'un faible nombre (de l'ordre de quelques dizaines) d'exemples. Cette propriété répond également au besoin applicatif de pouvoir créer (et sélectionner) de nouvelles réponses prototypiques (de nouvelles catégories) à partir d'un petit ensemble de courriels.

De surcroît, en dépit de performances relativement médiocres dans l'identification de la meilleure réponse (catégorie), cet apprenti s'est montré très performant pour positionner la classe correcte parmi les cinq choix proposés à l'utilisateur, surpassant pour cette mesure les meilleurs systèmes de catégorisation (SVM, k-plus-proches-voisins). Ces constatations ont été vérifiées sur plusieurs bases de données textuelles (forums de discusions, corpus de pourriels, mail center).

3.2.2 Classifier pour catégoriser

Des travaux complémentaires ont conduit à approfondir ce modèle de façon à (i) pouvoir distinguer des catégories de documents ne correspondant pas directement à des différences de contenu thématique et à (ii) prendre en compte la dérive temporelle des catégories. Ces deux problématiques sont particulièrement importantes pour l'application de *mail center*, dans la mesure où d'une part les catégories (ici les réponses prototypiques) sont proposées par les gestionnaires du système et peuvent être thématiquement très ressemblantes et où, d'autre part, la forme des requêtes peut présenter une forte variation au cours du temps.

Les diverses améliorations envisagées tournent autour de la question de la gestion temporelle d'une population de prototypes, chaque classe pouvant, être représentée par un ou plusieurs prototypes, qui correspondent

[7] Il suffit pour cela de recalculer la position des prototypes à chaque fois que de nouveaux exemples étiquetés sont rendus disponibles.

alors à des sous-thèmes. La construction non-supervisée des sous-thèmes pour des tâches de catégorisation de documents est également exploitée dans [Oh et al., 2001, Vilalta and Rish, 2003].

Vinot and Yvon [2003] propose et compare plusieurs algorithmes classiques (k-means) et originaux (reposant sur des techniques de classification hiérarchique) pour construire, de manière non-supervisée, ces populations de prototypes, une préoccupation étant de ne pas multiplier inutilement les prototypes utilisés par l'apprenti. La contribution principale de ce travail est de montrer qu'en guidant le clustering par des informations de supervision, il est possible de construire des sous-thèmes qui permettent d'améliorer de manière significative les performances du système de catégorisation, *tout en augmentant de manière minimale l'ensemble des prototypes construits*.

Pour entrer un peu dans le détail, notons p une sous-classe de documents appartenant à la catégorie $c(p)$. L'idée générale des algorithmes descendants (les algorithmes ascendants exploitent des idées similaires) est de ne proposer une bi-partition de p (en p_1 et p_2) que s'il existe une sous-classe p' appartenant à la catégorie $c(p') \neq c(p)$ qui est à la fois plus proche de p_1 que p_2 et plus proche de p_2 que p_1. Ce critère modélise le fait qu'une partition de p n'est nécessaire que s'il existe une autre sous-classe p' dont la proximité avec p risque d'entraîner des confusions entre les deux catégories $c(p)$ et $c(p')$. L'algorithme proposé dans [Vinot and Yvon, 2003] consiste à explorer les partitions possibles des catégories initiales, en n'acceptant une partition que si le critère précédent est vérifié. Si c'est le cas, la partition est effectuée et donne lieu à de nouvelles sous-classes, qui pourront à leur tour être partitionnées. Un critère d'arrêt supplémentaire permet d'éviter la création de sous-classes d'effectif trop faible.

Les expérimentations décrites dans cet article montrent que nos algorithmes de partionnement sont capables de s'adapter à la complexité de la tâche de catégorisation : des sous-classes ne sont effectivement créées que si les catégories d'origine se recouvrent fortement. En termes de performances, cette variante de l'algorithme de Rocchio (utilisant plusieurs prototypes par classes) permet d'obtenir des performances en généralisation significativement meilleures que celles obtenues avec la version originale de l'algorithme.

3.2.3 Conclusion

Ce travail est principalement un travail expérimental autour de la réalisation d'une application opérationnelle, dont les principales idées ont fait l'objet d'une implantation en vraie grandeur dans un système exploité commercialement[8]. Il apporte également, de manière détournée, des éléments empiriques de réponse à deux questions importantes (voir les discussions de la section 1.3) :

– la question du choix entre apprentis discriminants et génératifs : en dépit de leur moindre performance pour identifier la meilleure classe, les modèles génératifs peuvent apporter une plus-value réelle dans un cadre applicatif semi-supervisé. Dans la mesure, en effet, où l'algorithmes de Rocchio construit un modèle génératif des classes, il permet également de modéliser leurs similarités, permettant de trouver toutes les classes potentiellement intéressantes pour un document. Par contraste, les techniques discriminantes ne modélisent que la frontière entre documents et se montrent, dans notre contexte, moins performantes. Le débat entre les deux stratégies d'apprentissage ne peut donc pas uniquement être tranché en théorie, mais doit faire l'objet d'une évaluation dans chaque contexte applicatif envisagé.
– la question de la sélection automatique de la dimension du modèle (correspondant ici à la détermination du nombre correct de prototypes) : l'augmentation du nombre de sous-classes pour une catégorie donnée correspond à une augmentation de la taille et de la complexité de la modélisation des données. En utilisant un critère qui intègre des informations de supervision, l'algorithme de classification présenté ci-dessus adapte automatiquement la dimension du modèle à la difficulté de la tâche.

[8] http://www.akio-software.com.

3.3 Un modèle génératif pour la classification de documents

Cette section présente les grandes lignes d'un travail[9] portant sur l'étude de modèles probabilistes pour la classification non-supervisée de documents. L'existence de grandes collections de données textuelles rend nécessaire le développement d'outils d'aide à la structuration, à la navigation et à la visualisation de ces grandes masses de données. La plus simple des structurations consiste à former des classes (*clusters*) thématiquement homogènes de documents. De tels regroupements ont une utilité aussi bien dans une démarche d'analyse exploratoire de la collection, pour des tâches de recherche documentaire, de décision (catégorisation supervisée de documents[10], mais également pour identifier des sous-corpus homogènes pour des tâches de modélisation linguistique.

De multiples techniques de classification non-supervisée ont été recensées et étudiées y compris pour des données textuelles (en partie les travaux déjà ancien en lexicométrie, stylométrie, analyse de questionnaires, etc. [Lebart and Salem, 1988]). La numérisation de larges fonds documentaires, la diffusion des outils de recherche documentaire ont suscité un fort regain d'intérêt envers ces méthodes, conduisant à de nouvelles propositions de modèles probabilistes pour la classification non-supervisée [Nigam et al., 2000, Hofmann, 2001, Griffiths and Steyvers, 2002, Blei et al., 2002, Buntine and Jakulin, 2004]. De manière indirecte, ces techniques permettent également de construire des mesures de similarité entre mots, utilisables dans de nombreux contextes ou encore d'opérer une réduction de l'espace dans lequel les documents sont représentés, dont peuvent bénéficier d'autres applications de gestion documentaire.

Dans ce contexte, la contribution principale de ce travail est une analyse expérimentale très minutieuse du comportement d'un modèle probabiliste relativement simple, dans lequel les documents sont modélisés par un mélange de lois multinomiales. Les principales questions étudiées portent sur les problèmes liés à l'évaluation des regroupements construits, à la variabilité des estimations en très grande dimension et à l'étude comparative de divers algorithmes d'estimation des paramètres du modèle (EM et des variantes, méthodes à base de simulations).

3.3.1 Le modèle

Le modèle étudié reprend une proposition originellement formulée par [Nigam et al., 2000, Clérot et al., 2004], dans lequel la probabilité d'un document est modélisée par un mélange de n_C lois multinomiales. Le modèle de génération d'un document d de longueur $l(d)$ est le suivant :
– tirer une classe c selon une distribution discrète de paramètre $\alpha = \{P(c), c \in \mathcal{C}\}$;
– tirer $l(d)$ mots selon une loi multinomiale de paramètres $\{\beta_c = P(w|c), w \in \mathcal{V}\}$
Notant, comme précédemment, θ le vecteur de paramètres, la probabilité d'un document s'exprime par :

$$
\begin{aligned}
P(d;\theta) &= P(\theta)P(w_1,\ldots,w_{l(d)}|\theta) && (3.4)\\
&= P(\theta)\sum_{c=1}^{n_C} P(c|\theta)P(w_1,\ldots,w_{l(d)}|c,\theta) && (3.5)\\
&= P(\theta)\sum_{c=1}^{n_C} P(c|\theta)l(d)!\prod_{w}\frac{P(w|c,\theta)^{\mathrm{N}(w,d)}}{\mathrm{N}(w,d)!} && (3.6)\\
&= P(\theta)\frac{l(d)!}{\prod_{w=1}\mathrm{N}(w,d)!}\sum_{c=1}^{C} P(c|\theta)\prod_{w=1}^{n_V} P((w|c,\theta)^{\mathrm{N}(w,d)}) && (3.7)
\end{aligned}
$$

[9] Ce travail a été conduit en collaboration avec L. Rigouste et O. Cappé, dans le cadre d'un contrat de recherche avec l'équipe de F. Clérot à France Télécom R&D.

[10] cf. la discussion de la section précédente, voir aussi [Rigouste et al., à paraître, 2006] pour une application à la segmentation de textes.

Pour lisser les paramètres θ du modèle (le vecteur α, la matrice β $(n_V \times n_C)$), nous supposons, pour chacun, une distribution *a priori* de Dirichlet.

À nombre de classes fixées, il est possible de mettre en œuvre un algorithme de type EM pour estimer les différents paramètres du modèle (on se reportera pour les détails techniques à [Rigouste et al., 2005a]). Au terme de l'apprentissage, la classe du document d est celle qui a la plus forte probabilité *a posteriori* :

$$c^* = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(d|c)P(c) \tag{3.8}$$

3.3.2 Quelques observations empiriques

De multiples expérimentations ont été menées pour évaluer et analyser les performances de ce modèle. Pour faciliter l'évaluation, ces expérimentations ont initialement utilisé un corpus équilibré, extrait du corpus Reuters en sélectionnant 5 catégories très différentes, puis en choisissant 1000 documents par catégorie. Pour ce corpus, il existe donc une catégorisation de référence, fournie par les étiquettes de classe.

Les premiers enseignements de ces expérimentations sont tirés dans [Rigouste et al., 2005a]. Pour en synthétiser les principaux, notons que :

– Si la convergence de EM est rapide, elle conduit le plus souvent à atteindre un des multiples optima locaux de la fonction de vraisemblance. En conséquence, *le choix d'une bonne initialisation pour le modèle se révèle crucial*. Les différentes stratégies d'optimisation aléatoire évaluées conduisent à des modèles, qui sont, en termes de vraisemblance, beaucoup moins performants que le modèle obtenu en choisissant une initialisation idéale à partir des étiquettes de catégorie.
– En conséquence, quelle que soit la stratégie d'initialisation choisie, les regroupements construits restent très éloignés du regroupement idéal que constituent les catégories Reuters. En revanche, l'apport d'une information de supervision très réduite, exploitant les étiquettes d'une dizaine de documents par classe, permet de reconstituer des regroupements qui sont très proches de la catégorisation initiale.
– La très grande dimensionnalité du modèle (le vocabulaire d'indexation complet comporte plusieurs dizaines de milliers de termes) conduit à des affectations de classe qui sont quasi-déterministes : pour la majorité des documents, la classe optimale a une probabilité très proche de 1. Cette constatation nous a conduit à vérifier qu'une approximation déterministe de l'algorithme EM (consistant à remplacer l'étape E par des affectations déterministes dans la classe la plus probable) aboutit à des résultats identiques à ceux obtenus avec la version probabiliste.
– Travailler dans des espaces de moindre dimension, obtenus en restreignant le vocabulaire d'indexation aux termes les plus fréquents, facilite la convergence de l'algorithme vers un meilleur[11] optimum local et réduit nettement la variabilité des performances. Néanmoins, les performances globales sont dégradées par rapport à un modèle qui inclut tous les termes d'indexation.

3.3.3 Améliorer l'estimation

Ces observations nous ont conduit à formuler et à examiner diverses propositions visant à améliorer les estimateurs obtenus, tout en en réduisant la variabilité.

Nous avons en premier lieu proposé une méthode incrémentale d'estimation [Rigouste et al., 2005b], dans laquelle le vocabulaire est introduit de manière progressive. Cette méthode utilise le fait que pour l'algorithme EM, il est équivalent de disposer des paramètres ou des probabilités *a posteriori* d'appartenance aux classes. En s'intéressant à ces dernières, il est possible de mettre en œuvre des approches dans lesquelles le nombre de paramètres varie. Le principe de notre méthode est le suivant : nous initialisons l'algorithme

[11] Meilleur au sens où il correspond à des valeurs de la vraisemblance plus proches de celles que l'on obtient avec une initialisation supervisée.

avec un modèle estimé sur un vocabulaire d'indexation très réduit (quelques dizaines ou quelques centaines de mots) ; chaque accroissement ultérieur du vocabulaire donne lieu à un nouveau round d'estimation (EM), *qui est initialisé à partir des probabilités a posteriori calculées à l'étape précédente*.

Cette heuristique nous a permis d'obtenir des *clusters* à la fois plus stables et plus ressemblants aux catégories d'origine que ceux obtenus par EM. Elle s'avère toutefois plus exigeante en temps de calcul (l'algorithme EM est utilisé de manière répétée) et demande de régler un paramètre supplémentaire, qui est le rythme d'accroissement du vocabulaire. Ces deux inconvénients nous ont conduit à envisager d'autres méthodes d'estimation, en particulier des méthodes de simulation. En nous appuyant sur une remarque de [Griffiths and Steyvers, 2002], qui détaille la mise en œuvre de techniques de MCMC (*Monte-Carlo Markov Chains*) pour le modèle LDA, nous avons proposé et évalué, dans [Rigouste et al., 2006a], une technique analogue pour le modèle de mélange de multinomiales. Les résultats obtenus montrent que pour ce modèle, les méthodes de simulation permettent effectivement d'obtenir des performances meilleures que EM. La variabilité des estimateurs obtenus reste une faiblesse préoccupante de ce modèle (comme d'ailleurs d'autres modèles apparentés tels que LDA [Rigouste et al., 2006b]), du moins dans une optique d'analyse exploratoire de corpus.

3.3.4 Un bilan et quelques perspectives

Cette étude nous a permis d'analyser en profondeur le comportement d'un modèle probabiliste pour la classification de textes, conduisant à identifier assez précisément ses avantages (paramètres facilement interprétables, capacité de généralisation à de nouveaux documents) et les limites (variabilité des estimateurs en grande dimension, décisions d'appartenance quasi-déterministes, etc.), certaines de ces conclusions valant également pour d'autres modèles probabilistes apparentés. Il reste encore beaucoup à faire sur ce modèle, et les travaux en cours ou envisagés se développent dans plusieurs directions.

Dans le modèle multinomial, tous les mots d'un document sont tirés dans une même distribution, qui est déterminée par le choix du thème. Cette hypothèse est probablement trop forte, dans la mesure, par exemple, où certains mots fréquents ont une forte probabilité pour tous les thèmes. Ce modèle conduit, on l'a vu, à des affectations de classe qui sont quasi-déterministes. Les modèles alternatifs, de type LDA [Blei et al., 2002], font une hypothèse différente, qui consiste à choisir un thème pour chaque mot du document, à partir d'une distribution fixée une fois par document. En conséquence, il semble que les regroupements obtenus soient réellement probabilistes : chaque document effectivement est distribué sur plusieurs thèmes. Il est donc intéressant de chercher à construire des modèles de complexité intermédiaire entre ces deux situations, dans lesquels chaque document est potentiellement multi-thématique, sans pour autant laisser la possibilité de changer de thème à chaque mot.

Dans le modèle multinomial, comme dans d'autres modèles probabilistes apparentés, il est nécessaire de fixer *a priori* le nombre de composantes du mélange. Dans une optique d'analyse exploratoire, devoir déterminer à l'avance le nombre de thèmes est irréaliste. Pourtant, un mauvais choix du nombre de thèmes peut conduire à des regroupements hasardeux et instables. Une solution heuristique consiste à effectuer des tests en faisant varier la taille du modèle [Griffiths and Steyvers, 2004].

À la suite notamment de travaux comme ceux de [Vinokourov and Girolami, 2002, Blei et al., 2004], nous nous intéressons à deux extensions plus fondées théoriquement : d'une part, l'utilisation de techniques de sélection de modèle ; d'autre part la conception d'une version hiérarchique du modèle multinomial, qui permettrait de déterminer *a posteriori* le nombre correct de clusters. Des expériences préliminaires sont présentées dans [Rigouste et al., 2005b].

Une seconde voie d'amélioration de ce modèle concerne plus directement les questions liées au passage à l'échelle des algorithmes d'estimation. La mise en œuvre d'EM, comme celle de méthodes de simulation, implique typiquement de parcourir plusieurs fois l'intégralité du corpus d'apprentissage. Si cette stratégie est

possible pour des corpus de dimension réduite, elle devient difficilement tenable dès lors qu'on s'intéresse à des collections plus grandes, contenant des centaines de milliers, voire des millions de documents (ou plus) ; ou encore si l'on travaille avec des ressources (temps ou mémoire) restreintes. Ces considérations nous ont conduit à nous intéresser à des variantes en ligne (ou récursives) de l'algorithme EM, qui n'examinent qu'une seule fois chacun des exemples de la base d'apprentissage.

3.4 Conclusion

Il est troublant, parfois aussi frustrant, de constater, qu'en matière de modélisation statistique du langage, comme en matière de fouille de textes, le paysage reste dominé par des modèles (k-grammes d'un côté, « sac de mots » de l'autre) qui partagent de nombreuses caractéristiques. Linguistiquement frustres, ces modèles se fondent sur une représentation surfaciste des textes et exploitent des hypothèses très discutables (conception simpliste de la notion de mot, localité des dépendances entre mots pour les k-grammes, indépendance conditionnelle des termes sachant la classe pour la catégorisation de documents). Leur efficacité, indiscutable, est à chercher dans leur dimensionnalité : dans les deux cas, il s'agit de manipuler des lois de probabilités discrètes sur un ensemble d'évènements contenant plusieurs dizaines de milliers d'éléments, dont l'estimation est pourtant très problématique (cf. la discussion de la section 1.3.2).

En dépit de ces travers unanimement reconnus, ces modèles restent extrêmement difficiles à améliorer. À cet égard, l'examen du bilan de quinze ans de travaux en modélisation du langage est tristement révélateur : les systèmes de l'état de l'art continuent de s'appuyer sur des modèles k-gramme et la seule méthode garantissant à coup sûr une amélioration des performances du reconnaisseur consiste à augmenter le volume des données d'entraînement et la taille des modèles. Ces constatations sont d'autant plus insatisfaisantes qu'elles semblent presque universellement valides, y compris pour des langues dont la morphologie et/ou la syntaxe sont très différentes de celle de l'anglais.

J'ai déjà déjà eu l'occasion, à la section 1.3.4, de dresser quelques perspectives générales pour le domaine de l'apprentissage statistique à partir de corpus textuels, dont beaucoup s'appuient sur ces observations. Aussi me contenterai-je, pour conclure ce chapitre d'indiquer quelques-uns des thèmes sur lesquels je compte faire porter mes propres travaux.

Fouille de textes Les outils de fouille de textes (catégorisation, clustering, segmentation) font aujourd'hui partie des technologies de base pour de nombreuses applications : recherche d'information, question/réponse, etc., qui traitent des documents très différents en longueur (variant du livre à la phrase) ou en structure (textes bruts, documents HTML, etc). L'effort doit être poursuivi pour améliorer ces techniques, en particulier pour mieux caractériser le champ d'application de telle ou telle approche. Un effort tout particulier, pour les textes courts, doit continuer de se porter sur les modèles capables d'extraire ou d'exploiter des représentations plus riches que le « sac de mots », intégrant en particulier des relations de dépendance syntaxiques, des informations terminologiques ou sémantiques, des informations sur la structure interne du document, des méta-données, etc. Ces informations sont aujourd'hui quasi-systématiquement disponibles, ou peuvent être calculées, du moins pour les langues les plus étudiées, avec une fiabilité grandissante.

Modélisation du langage Concernant la modélisation statistique du langage, j'envisage d'une part de persévérer dans l'étude de modèles intégrant des formes minimales de connaissances linguistiques (en particulier morphologiques) pour améliorer en particulier le lissage. Il me semble que le discrédit de ces méthodes est pour une large part dû à leur évaluation sur des corpus de langue anglaise, langue qui n'est sans doute pas celle auxquelles elles se prêtent le mieux.

L'autre question à travailler est celle de l'adaptation thématique dynamique des modèles de langage : je l'ai

mentionné à la section 1.3.1, l'approche la plus courante en modélisation du langage consiste à mélanger de manière heuristique des corpus variés pour construire un modèle unique, qui est utilisé pendant l'étape de décodage acoustico-phonétique. Cette approche, dont l'efficacité est indéniable, conduit toutefois à des systèmes de reconnaissance d'une complexité gigantesque. Il me semble aujourd'hui toujours nécessaire de poursuivre les réflexions sur l'adaptation dynamique au thème des modélisation du langage, afin de dépasser les propositions de e.g. [Kuhn and DeMori, 1990, Seymore and Rosenfeld, 1997, Florian and Yarowsky, 1999], notamment en s'inspirant des propositions récentes de [Griffiths et al., 2005].

Bibliographie

Ion Androutsopoulos, John Koutsias, Konstandinos V. Chandrinos, and Constantine D. Spyropoulos. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 160–167. ACM Press, 2000.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA, 2002. MIT Press.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, Vancouver, Canada, 2004.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proccedings of 11th annual conference on Computational Learning Theory (COLT 98)*, pages 92–100, Madison, WI, 1998.

Sergey Brin. Extracting patterns and relations from the World-Wide Web. In *Proceedings of the International Workshop on the Web and Databases (WebDB '98)*, 1998.

Wray Buntine and Aleks Jakulin. Applying discrete PCA in data analysis. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04)*, pages 59–66. AUAI Press 2004, 2004.

Fabrice Clérot, Olivier Collin, Olivier Cappé, and Eric Moulines. Le modèle "monomaniaque" : un modèle statistique simple pour l'analyse exploratoire d'un corpus de textes. In *Colloque International sur la Fouille de Texte (CIFT'04)*, La Rochelle, France, 2004.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6) :391–407, 1990.

Florence Duclaye. *Apprentissage automatique de relations d'équivalence sémantique à partir du World Wide Web*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris, France, 2004.

Florence Duclaye, Pascal Filoche, Jerzy Sitko, and Olivier Collin. A Polish question/answering system for business information. In *Fifth international conference on business information systems (BIS'05)*, Poznan (Poland), 2002a.

Florence Duclaye, François Yvon, and Olivier Collin. Using the web as a linguistic resource for learning reformulations automatically. In *Third international conference on language resources and evaluation (LREC'02)*, Las Palmas (Spain), 2002b.

Florence Duclaye, François Yvon, and Olivier Collin. Apprentissage automatique de paraphrases. In *Traitement Automatique des Langues Naturelles (TALN)*, Batz sur mer (France), 2003a.

Florence Duclaye, François Yvon, and Olivier Collin. Learning paraphrases to improve a question-answering system. In *EACL 03 Workshop on 'Question Answering Systems'*, Budapest (Hongrie), 2003b.

Florence Duclaye, François Yvon, and Olivier Collin. Unsupervised and incremental acquisition of a thematic corpus from the www. In *IEEE Natural Language Processing and Knowledge Engineering*, Beijing(China), 2003c.

Radu Florian and David Yarowsky. Dynamic nonlocal language modeling via hierarchical topic-based adaptation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 167–174, College Park, MD, 1999.

Antoine Ghaoui, François Yvon, Chafik Mokbel, and Gérard Chollet. On the use of morphological constraints in N-gram statistical language models. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, 2005.

Thomas L. Griffiths and Mark Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, Fairfax, VA, 2002.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (supl 1) :5228–5235, 2004.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua Tenenbaum. Integratic topics and syntax. In *Proceedings of NIPS*17*, Vancouver, CA, 2005.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, pages 539–545, 1992.

Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1) :177–196, 2001.

Thomas Hofmann and Jan Puzicha. Statistical models for co-occurence data. Technical Report AI Memo No 1625, AI Lab., MIT, 1998.

Thorsten Joachims. Text categorization with Support Vector Machines : Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML 98)*, pages 137–142, 1998.

Adam Kilgariff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3) :333–347, 2003.

Aleksandra Krul, Géraldine Damnati, Thierry Moudenc, and François Yvon. Corpus design based on the kullback-leibler divergence for text-to-speech synthesis applications. In *Proceedings of ICLSP/InterSpeech 2006*, Carnegie-Mellon, PA, 2006.

Roland Kuhn and Renato DeMori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6) :570–583, 1990.

Ludovic Lebart and André Salem. *Analyse statistique des données textuelles*. Dunod, Paris, France, 1988.

David D. Lewis. Naive (Bayes) at forty : The independence assumption in Information Retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15, 1998.

Dekan Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL'98*, volume 2, pages 768–774, Montréal, Canada, 1998.

Andrew K. McCallum and Kamal Nigam. Employing EM in pool-based active learning for text classification. In Jude W. Shavlik, editor, *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pages 350–358, Madison, US, 1998. Morgan Kaufmann Publishers, San Francisco, US.

Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3) :103–134, 2000.

J-H Oh, K-S Lee, D-S Chang, C. Won Seo, and K-S Choi. TREC-10 experiments at KAIST : Batch filtering and Question Answering. In *Proceedings of The Tenth Text REtrieval Conference (TREC-10)*, pages 347–354, 2001.

Bo Pang and Lillian Lee. Seeing stars : Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Annual Meeting of the North-American Chapter of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005.

Bo Pang, Lillian Lee, and Shivakumar Vaithyannathan. Thumbs up ? sentiment classification using Machine Learning techniques. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2002*, 2002.

Loïs Rigouste, Olivier Cappé, and François Yvon. Evaluation of a probabilistic method for unsupervised text clustering. In *Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, Brest, France, 2005a.

Loïs Rigouste, Olivier Cappé, and François Yvon. Inference for probabilistic unsupervised text clustering. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, France, 2005b.

Loïs Rigouste, Olivier Cappé, and François Yvon. Inference and evaluation of the multinomial mixture model for text clustering. Technical Report 2006D004, Télécom Paris, 2006a. URL http://www.tsi.enst.fr:8080/cgi-bin/valipub_download.cgi?dId=11.

Loïs Rigouste, Olivier Cappé, and François Yvon. Quelques observations sur le modèle LDA. In Jean-Marie Viprey, editor, *Actes des IXe JADT*, pages 819–830, Besançon, 2006b.

Loïs Rigouste, Olivier Cappé, and François Yvon. Modèles multi-thématiques Markoviens pour la segmentation de textes. *RNTI : revue des nouvelles technologies de l'information*, à paraître, 2006.

Ellen Riloff. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence, AAAI 99*, pages 474–479, Orlando, FL, 1999.

Joseph J. Rocchio. *The SMART Retrieval System : Experiments in Automatic Document Processing*, chapter 14, Relevance Feedback in Information Retrieval, pages 313–323. Gerard Salton (editor), Prentice-Hall Inc., New Jersey, 1971.

Mats Rooth. Two-dimensional clusters in grammatical relations. In *Proceedings of the symposium on representation and acquisition of lexical knowledge : polysemy, ambiguity and generativity*, Standford, CA, 1995.

Robert E. Schapire and Yoram Singer. BoosTexter : A boosting system for text classification. *Machine Learning*, 39(2/3) :135–168, 2000.

Robert E. Schapire, Yoram Singer, and Amit Singhal. Boosting and rocchio applied to text filtering. In W. Bruce Croft, Alistair Moffat, Cornelis J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 215–223, Melbourne, AU, 1998. ACM Press, New York, US.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1) : 1–47, 2002.

Kristie Seymore and Ronald Rosenfeld. Using story topics for language model adaptation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1987–1990, Rhodes, Greece, 1997.

Peter D. Turney. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3) : 251–278, 2005.

Peter D. Turney. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling/ACL-06)*, pages 313–320, Sidney, Australia, 2006.

Ricardo Vilalta and Irina Rish. Decomposition of classes via clustering to explain and improve Naïve Bayes. In Springer, editor, *Proceedings of the 14th European Conference on Machine Learning (ECML)*, pages 444–455, Cavtat-Dubrovnik, 2003.

Alexei Vinokourov and Mark Girolami. A probabilistic framework for the hierarchic organisation and classification of document collections. *Journal of Intelligent Information Systems*, 18(2/3) :153–172, 2002.

Romain Vinot. Classification automatique de textes dans des catégories non thématiques. Thèse de l'Ecole Nationale Supérieure des Télécommunications, Paris, France, 2004.

Romain Vinot and François Yvon. Semi-automatic response in a Mail Center. In *ASMDA 2001, 10th International Symposium on Applied Stochastic Models and Data Analysis*, pages 992–997, Compiègne, France, 2001.

Romain Vinot and François Yvon. Quand simplicité rime avec efficacité : Analyse d'un catégoriseur de textes. In *Colloque International sur la Fouille de Texte (CIFT'02)*, pages 17–26, Hammamet, Tunisie, 2002.

Romain Vinot and François Yvon. Improving Rocchio with weakly supervised clustering. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 456–467, Cavtat-Drubrovnick, Croatie, 2003.

Romain Vinot, Natalia Grabar, and Mathieu Valette. Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet. In *Actes de la 10ème conférence sur le Traitement Automatique du Langage Naturel (TALN)*, pages 275–284, 2003.

Yiming Yang. An evaluation of statistical approaches to text categorization. Technical Report CMU-CS-97-127, Carnegie Mellon University, 1997.

Chapitre 4

Contributions au traitement de la parole

Ce chapitre dresse une synthèse de mes travaux récents en traitement automatique de la parole, en distinguant de manière très classique ceux qui portent sur la synthèse vocale et notamment sur les questions de translittération orthographique-phonétique (section 4.1) ; de ceux qui portent sur la reconnaissance vocale et en particulier sur la constitution de ressources dites « linguistiques » (dictionnaires et modèles de langage) et sur leur intégration dans les systèmes de reconnaissance (section 4.2).

Mes contributions les plus récentes portent principalement sur la reconnaissance vocale, notamment sur la construction, l'optimisation et l'utilisation de transducteurs pondérés en reconnaissance vocale (voir, parmi les articles joints, les références [Yvon et al., 2004a, Shiuan-Sung and Yvon, 2005]).

4.1 Translittération orthographique-phonétique et synthèse vocale

Mes activités[1] en synthèse de la parole à partir du texte ont porté presqu'exclusivement sur les questions de translittération orthographique-phonétique (voir la section A.0.7 pour une description de la tâche), dans la continuation directe de mes travaux de thèse. Faute d'avoir réalisé leur intégration dans un système de synthèse complet, ces développements sont restés d'une portée modeste et cette activité est aujourd'hui quasiment arrêtée, hormis pour ce qui concerne l'apprentissage de la prononciation, qui reste un banc d'essai difficile pour les outils d'apprentissage automatique[2].

Ces activités s'organisent sur trois axes principaux :
– le développement de systèmes à base de règles ;
– le développement de systèmes utilisant des techniques d'apprentissage ;
– la constitution et validation de dictionnaires de prononciation pour la synthèse et la reconnaissance vocale ;

Le second de ces axes constitue un prolongement direct de mon travail de thèse [Yvon, 1996c, Deligne et al., 1996, Yvon, 1996b] ; une large place ayant étant accordée à ce type de travaux à la section 2.3.1, ils ne seront pas développés ici plus avant.

Le dernier de ces axes ne sera pas non plus développé plus avant, la contribution la plus intéressante restant, hormis naturellement le développement de ces ressources et leur mise à disposition, une réflexion sur la

[1] Ces travaux ont bénéficié d'échanges très riches avec V. Aubergé (ICP), F. Béchet (LIA), P. de Mareüil et C. d'Alessandro (LIMSI). Travailler avec eux fût (et reste) un plaisir : qu'ils en soient remerciés. Y. Régis-Gianas a réalisé un remarquable travail d'implantation sur la transcription orthographique phonétique et mérite largement d'être associé à ces remerciements. V. Beaudouin m'a poussé à adapter mes outils de transcription au traitement de l'alexandrin classique, initiant une collaboration qui je l'espère durera encore longtemps.

[2] Voir http://www.pascal-network.org/Challenges/PRONALSYL/.

représentation des variantes de prononciation dans [Boula de Mareuïl et al., 1998, Boula de Mareüil et al., 2000, Yvon et al., 2000].

Il m'a en revanche semblé pertinent de revenir ici sur une question en apparence bien documentée, celle des systèmes à base de règles pour la transcription. Bien que cette problématique ait quasiment entièrement disparu des préoccupations de la communauté, il m'apparaît qu'avec l'avènement d'une nouvelle génération de systèmes de synthèse plus naturels, les besoins de produire une parole plus expressive et variée vont susciter de nouveaux travaux autour des variantes de prononciation et des moyens de les engendrer, en particulier dans le cadre des évolutions des systèmes à base de règles, qui restent de loin les plus utilisés en synthèse vocale.

La section 4.1.1 présente donc successivement deux évolutions de mon système à base de règles, la plus récente [Régis-Gianas and Yvon, 2004] me semblant constituer une réponse moderne à la question de l'élaboration de règles de prononciation (à comparer par exemple avec l'archaïsme relatif de l'état de l'art en la matière, représenté par exemple par le système Festival [3]).

Quelques perspectives plus générales sur le problème de la transcription sont ensuite discutées à la section 4.1.2.

4.1.1 Les systèmes à base de règles

Dans cette section, la problématique de la transcription orthographique-phonétique à base de règles est rapidement rappelée, à la suite de quoi les deux systèmes de transcription que j'ai développés sont présentés.

La récriture contextuelle

La conception d'un système de conversion graphème-phonème à base de règles repose essentiellement sur une technologie simple et formellement bien décrite, la *récriture contextuelle* (voir e.g. [Kaplan and Kay, 1994, Mohri and Sproat, 1996]).

Dans le contexte de la conversion graphème-phonème, une règle de récriture contextuelle exprime la transformation d'un motif (orthographique) en sa contre-partie phonétique, conditionnée par le contexte d'occurrence (gauche et/ou droit) du motif. Une récriture est dénotée formellement par :

$$\phi\,[a]\,\psi \;\rightarrow\; \lambda\,[b] \tag{4.1}$$

qui exprime la récriture du motif a en b dans un contexte décrit par des expressions rationnelles ϕ et ψ. Lorsque ce formalisme est mis en œuvre dans un parcours gauche-droite de l'entrée orthographique, il est également possible de préciser le contexte phonétique gauche λ de b en sortie.

D'un point de vue opérationnel, une récriture est interprétée comme suit : si à la position courante de l'entrée, il existe un sous-mot capturé par a, précédé (respectivement suivi) d'un segment capturé par ϕ (resp. ψ), si le suffixe de la sortie courante est capturé par λ alors (i) on concatène b à la sortie courante, (ii) le pointeur de l'entrée courante est déplacé à droite de a. Cette opération peut être calculée efficacement par un transducteur fini.

Un système de transcription est constitué d'un ensemble de règles de récriture de la forme (4.1) décrivant les correspondances entre graphèmes et phonèmes. Pour le français, une description de ces règles est donnée par exemple dans [Fouché, 1959, Catach, 1984]. Les exceptions se prêtent à une formalisation du même type, les récritures correspondantes spécifiant des contextes gauche et droit dénotant des frontières de mot.

[3] festvox.org/festival

La conversion graphème-phonème pour la synthèse vocale

Dans la réalité, l'opération de transcription orthographique-phonétique est passablement plus complexe, puisque la seule connaissance de la chaîne orthographique à transcrire ne suffit pas à déterminer de manière non ambiguë une prononciation : l'existence d'*homographes-hétérophones* impose d'introduire des éléments de contexte supplémentaires, tel qu'en particulier l'étiquette morpho-syntaxique du mot à prononcer. D'autres éléments de contexte potentiellement utiles en français concernent la structure morphologique de la forme à transcrire, la localisation du mot dans l'énoncé ou au sein du groupe prosodique (voir, par exemple, [Divay, 1990, de Mareüil, 1997, Béchet, 2001]).

Notons également que l'étape de transcription, si elle constitue le cœur du système de phonétisation, n'est pas le seul composant impliqué : il est ainsi préalablement nécessaire, de normaliser les textes (surtout lorsqu'ils sont très « bruités », par exemple des courriels), de réaliser un étiquetage morpho-syntaxique, une analyse morphologique (au moins grossière), d'identifier l'origine linguistique des mots, de calculer les endroits où une liaison peut être réalisée, de segmenter l'entrée en groupes prosodiques, etc. De même, la sortie phonétique doit être retravaillée (identification des syllabes phonétiques, localisation des accents, calcul des durées des syllabes, etc.) avant que l'on puisse procéder à l'étape de synthèse proprement dite [Dutoit, 1997].

Lors de ces différentes étapes de traitement, le formalisme des règles de récriture est un outil indispensable, qui permet d'exprimer des régularités aux différents niveaux d'analyse de la langue et d'implanter efficacement le calcul de ces régularités [Sproat, 1997].

Un phonétiseur pour le français

Dans mon travail de thèse [Yvon, 1996a], j'ai présenté un système complet, intégrant l'ensemble des traitements décrits ci-dessus pour la transcription de textes dans un contexte de synthèse vocale. Sans entrer dans les détails, la principale originalité de cet outil tient principalement à l'organisation du système de règles en trois modules principaux :

- un module de normalisation orthographique effectuant une conversion graphème-graphème et dédié à la gestion de toutes les exceptions ; ce module étant lui-même organisé en sous-modules dédiés aux différentes catégories d'entrée (mots du lexique commun ; sigles ; noms propres ; numéraux ; etc.)
- un module de conversion orthographique-phonétique restreint à des règles de portée très générale ;
- un module d'ajustement phonologique, modélisant les principales règles de la phonologie du français (nasalisation, traitement des e-muets, liaison, principales règles d'assimilation, etc.)

L'avantage d'une telle organisation, en plus de la maintenabilité globale du système, était qu'elle permettait, en jouant sur les règles impliquées dans le troisième module, de produire en sortie différents registres de phonétisation. En guise d'application de cette fonctionnalité, nous avons configuré un module de phonétisation spécifiquement dédié à la prononciation de l'alexandrin classique, et utilisé ce module pour transcrire l'intégralité des pièces en vers de Racine, Corneille et Molière (voir [Beaudouin and Yvon, 1996, 2004] pour une présentation de ces travaux et diverses applications à l'étude de la métrique et à l'attribution d'auteur).

Parmi les autres aspects remarquables de ce système, signalons qu'il s'exécute comme une surcouche de l'analyseur syntaxique Sylex [Constant, 1991] et qu'il accorde une attention toute particulière au problème difficile de la prononciation des extra-lexicaux, en particulier des noms propres [Vitale, 1991, Golding, 1991]. Dans ce phonétiseur, plus d'un millier de règles d'identification et de transcription sont en effet dédiées au traitement des noms propres et des sigles.

En ce qui concerne les noms propres, la transcription opère en deux étapes : dans un premier temps, le nom propre est affecté à une ou plusieurs grandes familles de langues, sur la base principalement de la reconnaissance de préfixes et suffixes caractéristiques. Selon les résultats de cette identification, le nom

propre est alors dirigé vers des règles de transcriptions spécifiques à chacune de ces familles de langues. En ce qui concerne les sigles, l'algorithme de prononciation reprend les principes généraux énoncés dans [Plénat, 1992, Yvon, 1994, 1996c].

Ce système de transcription a été évalué dans le cadre de la campagne d'évaluation des systèmes de synthèse vocale organisée sous l'égide de l'AUF. Les principaux résultats de cette évaluation, ainsi que les enseignements cette campagne quant à l'évalution des systèmes de transcription sont décrits dans [Yvon et al., 1998].

Dimi : récriture pour la synthèse et la reconnaissance

Le développement du système Dimi [Régis-Gianas and Yvon, 2004], entrepris en collaboration avec Yann Régis-Gianas, vise à proposer un environnement entièrement déclaratif pour le développement de règles de récriture, avec toujours comme objectif principal la construction d'un module de transcription orthographique-phonétique pour la synthèse vocale. Une contrainte supplémentaire est que le système puisse fonctionner de manière indépendante de l'analyseur syntaxique.

Dans le cadre de ce projet, plusieurs objectifs nouveaux sont donc mis en avant, qui le distinguent de la majorité des systèmes de l'état de l'art (en particulier en ce qui concerne les points b., c. et d.) :

a. l'intégralité du système de transcription doit pouvoir être décrit de manière déclarative, sans pour autant renoncer à l'efficacité du traitement ;
b. toutes les connaissances impliquées dans la translittération (dictionnaires, règles morphologiques, règles de transcriptions, règles phonologiques) sont représentées dans un formalisme homogène et cohérent ;
c. la syntaxe des règles doit s'accommoder des besoins particuliers de la synthèse et de la reconnaissance vocale. En plus de la gestion de contextes étendus, un problème récurrent de la transcription concerne le « bon moment » pour appliquer une règle. Ce problème est illustré par l'exemple des liaisons. Pour prédire une liaison, il faut d'une part, repérer une consonne latente, d'autre part vérifier le contexte syntaxique du mot courant ainsi que son contexte phonologique droit (voyelle ou consonne). Ce contexte étant encore inconnu au moment du repérage de la consonne de liaison (la transcription opère de gauche à droite), deux alternatives sont possibles : (i) essayer de décrire tous les contextes orthographiques droits qui peuvent donner lieu à une voyelle phonétique ou (ii) attendre d'avoir réalisé la transcription du contexte droit pour prendre une décision. L'option (ii) nous semble nettement préférable, puisqu'elle évite d'écrire deux fois les règles de prononciation des voyelles ; sa mise en œuvre nécessite toutefois un mécanisme pour propager de l'information (ici qu'une consonne latente a été détectée) entre des couches de traitement.
d. le système doit également gérer des cas de non-déterminisme (multiplicité des sorties), pour pouvoir être utilisé pour des applications de reconnaissance vocale ;
e. le système doit inclure un environnement complet de test, de déboguage et de traçage des règles, qui sont indispensables pour développer et maintenir des règles de transcription pour des langues telles que le français ;

Le respect de ces desiderata nous a conduit à redévelopper intégralement un langage pour exprimer les règles de transcription. En premier lieu, ce formalisme étend la syntaxe des règles décrite en (4.1) selon :

$$\phi\,[a]\,\psi \;\rightarrow\; \lambda\,[b]\,|\,-\;[P] \tag{4.2}$$

Cette nouvelle syntaxe introduit deux innovations. La première consiste à autoriser du non-déterminisme dans la partie droite d'une règle : dans la syntaxe de Dimi, le motif b peut en effet dénoter un langage fini arbitraire. Cette fonctionnalité est particulièrement utile pour décrire le lexique interne de Dimi[4], comme dans l'exemple suivant, qui exprime la possibilité d'effectuer ou non une diérièse[5] sur le mot *lié* :

[4] Dans ce formalisme, les entrées dictionnairiques sont traitées exactement comme des règles.

[5] Une *diérèse* correspond à la réalisation d'une séquence voyelle haute + voyelle sous la forme de deux voyelles « pleines », chacune constituant un noyau syllabique valide ; à l'inverse, la *synérèse* correspond à une réalisation dans laquelle la voyelle haute

```
[l i é] -> l (y|i) e |- Word
```

La seconde innovation consiste à contraindre l'application de la règle par la vérification d'une formule $[P]$ en logique des prédicats. Il est ainsi possible d'exprimer simplement qu'une règle concerne un mot complet (c'est le cas de la règle précédente), ne s'applique qu'en début ou en fin de mot, que pour une catégorie morpho-syntaxique donnée, etc. Les prédicats peuvent porter sur des fragments (intervalles) arbitraires des flux d'entrée et de sortie.

Ces fonctionnalités du langage sont illustrées par les quelques exemples de règles suivants, décrivant la prononciation de la lettre 'a' (la variable %V désigne l'ensemble des voyelles).

```
-- ex: taon
[a o n]    -> an |- EndOf_Word;

-- ex: aurore
[a u] r    -> o;

-- ex: laye, maye, braye
[a y e]    -> ai |- EndOf_Word;

-- ex: essayage, raya
[a y] %V   -> ai y;
```

De manière complémentaire, le langage permet d'accumuler (toujours sous forme de prédicats) des connaissances sur la partie du flux concernée par une règle. Il est ainsi possible de passer de règle en règle les informations accumulées au cours du traitement. Ainsi, par exemple, la règle suivante :

```
-- à la fin d'un mot, le 'b' n'est pas tronqué
[b] -> b |- EndOf_Word => NoTrunc;
```

exprime le fait qu'un 'b' final se prononce /b/, lequel /b/ devra échapper (dans un moment ultérieur du traitement, correspondant à l'application du module phonologique) à la règle de troncation des consonnes finales. Un exemple plus complexe d'utilisation est fourni par la gestion de la liaison évoqué ci-dessus. Dans Dimi, la liaison est traitée en trois étapes, selon le mécanisme suivant :

– dans un premier temps, les consonnes latentes correspondant principalement aux morphèmes flexionnels (terminaisons usuelles de la conjugaison verbale et de la flexion des noms et adjectifs) sont identifiées ; un prédicat exprimant la potentialité d'une liaison est alors associé à ces consonnes ;
– dans un second temps, l'environnement syntaxique de la liaison est évalué par application des règles usuelles de la liaison (liaison obligatoire entre un déterminant et le nom qu'il détermine, entre un pronom personnel sujet et le verbe, etc.)
– dans un troisième temps, l'environnement phonologique des consonnes latentes en position où une liaison est possible est enfin testé : si la consonne latente est suivie d'une voyelle, alors elle sera effectivement réalisée dans le flux de sortie.

La modularité et la déclarativité du système est garantie par l'organisation des règles en *modules* et *stratégies*, qui sont compilables et testables indépendamment les uns des autres. Un module est simplement un ensemble de règles partageant un trait commun : par exemple les règles qui identifient les marques flexionnelles, ou les règles qui concernent les sigles constituent un module. L'utilisation des modules est exprimée par des *stratégies*, qui décrivent comment les règles (d'un ou plusieurs modules) sont appliquées. Le langage de définition des stratégies définit les opérateurs suivants :

mute en glide, donnant lieu à une syllabe unique. Dans le langage courant, l'alternance diérèse/synérèse correspond à des variations de débit ; dans la poésie classique, elle a un impact sur le décompte syllabique et est encadrée par des règles très strictes.

– le choix déterministe (`<+`), qui applique une première stratégie et, en cas d'échec, applique la seconde. C'est le cas d'interaction entre règles le plus habituellement rencontré ;
– le choix non-déterministe (`+`), qui applique deux stratégies et conduit à la production de deux résultats ;
– le séquencement (`;`), qui applique une première stratégie sur le flux entier et, ensuite, la seconde sur le flux résultat. Cet opérateur peut être vu comme effectuant un *pipe-line* entre deux stratégies.

Ces fonctionnalités sont illustrées par le fragment de code suivant :

```
-- Choix déterministe
-- Les règles générales du francais pour la lettre 'a', constituant
-- le module ''G2P.letter_a, sont enchaînées de manière déterministe
strategy letter_a = (<+ #G2P.letter_a.*);

-- Choix non-déterministe: diverses prononciations d'un nom propre
-- avec plusieurs jeux de règles différents
strategy proper = (!French.proper + !English.proper + !Spanish.proper)
```

Une maquette pour la transcription du français intégrant une (petite) sous-partie du phonétiseur complet décrit à la section 4.1.1 a permis de tester ce nouvel environnement et d'en évaluer la pertinence. L'ensemble du système peut être téléchargé à l'adresse : `http://www.enst.fr/~yvon/Projets/dimi`.

4.1.2 Conclusion et perspectives

La question de la conversion graphème-phonème semble, dans le cadre des applications de synthèse vocale, une question résolue : des systèmes à base de règles couvrant de manière satisfaisante les principales régularités sont aujourd'hui disponibles pour le français [Béchet, 2001], comme pour l'anglais [Black and Taylor, 1997] ; de même que sont disponibles de larges dictionnaires phonétisés (pour le français BDLEX [Pérennou et al., 1991] ou ILPho [Yvon et al., 2000]). Les systèmes les plus performants atteignent, sur des textes journalistiques, des performances approchant 99% de phonèmes correctement prédits [Yvon et al., 1998].

Deux problèmes me semblent toutefois mal résolus. D'une part la prononciation des extra-lexicaux (en particulier des noms propres) reste une tâche difficile, pour laquelle les meilleurs systèmes peinent encore à obtenir des performances satisfaisantes (ainsi, en anglais, les taux d'erreurs par mot restent largement au delà des 20%). D'autre part, la modélisation des variantes de prononciation au niveau des phrases qui permettait, par exemple, de simuler différents registres d'élocution, est ignorée dans la majorité des systèmes existants, qui ne permettent d'engendrer qu'une seule prononciation.

Le relatif désintérêt pour ces problèmes provient pour partie, me semble-t-il, de la difficulté qu'il existe à évaluer correctement la qualité des systèmes de transcription : en synthèse vocale, plus encore qu'en reconnaissance vocale, l'existence de protocoles objectifs de mesures des performances est essentiel au développement des technologies. Deux modes d'évaluations des transcripteurs sont en effet couramment utilisés : la comparaison à une prononciation normative unique ; et l'évaluation de systèmes complets. Dans le premier protocole, se pose la question difficile de la détermination non plus d'une, mais d'un ensemble de prononciations de référence : c'est vrai pour le choix des transcriptions possibles d'une phrase, comme pour la définition des prononciations de références pour un nom propre. L'utilisation du second protocole demanderait de savoir mesurer l'impact des erreurs de prononciation sur la qualité globale du système. En synthèse, toutefois, l'impact de ces erreurs est difficile à séparer de celles des autres composants du systèmes, que ces erreurs soient commises préalablement ou postérieurement à la transcription elle-même.

Notons que la question de l'évaluation de la qualité des transcriptions se pose à l'identique en reconnaissance vocale et que les avis divergent sur les bénéfices d'une modélisation fine des variantes de prononciation (nous reviendrons sur ces questions à la section 4.2.3).

En ce qui concerne la synthèse vocale, il est probable toutefois que l'avènement d'une nouvelle génération de synthétiseurs par concaténation d'unités longues (voir e.g. [Black and Taylor, 1994, Campbell and Black, 1996, Donovan, 1996]), fondés sur l'exploitation de larges bases de données acoustiques conduira à terme, à reconsidérer la question de la prononciation. Un aspect important de ces sytèmes est qu'ils sont à même de s'accomoder de variantes de prononciation en entrée de l'algorithme de recherche de la meilleure séquence d'unités. Nul doute que cette possibilité, pour l'instant peu exploitée, fera à terme l'objet de nouvelles études concernant la génération de variantes.

4.2 Reconnaissance de la parole

Cette section présente une synthèse de mes travaux[6] sur la reconnaissance vocale. Ce domaine, que j'ai initialement abordé comme un débouché presque naturel des ressources linguistiques (dictionnaires et modèles de langage statistiques) sur lesquelles je travaillais, a progressivement occupé une part très importante de mon activité, du moins jusqu'à mon séjour au centre de recherche d'IBM (2001-2002). La poursuite d'une recherche active sur ces questions est aujourd'hui loin d'être une évidence, au vu notamment de la maturation rapide des technologies ainsi que des investissements importants en temps et en personnels qui sont nécessaires pour se maintenir au niveau de l'état de l'art.

En plus d'activités de développement de ressources et de systèmes, ma contribution principale me semble avoir porté sur les aspects algorithmiques liés à l'intégration, au sein des décodeurs de parole, des différentes ressources et particulier des ressources linguistiques. J'ai ainsi participé au développement de deux systèmes de reconnaissance vocale. Le système Sirocco en particulier témoigne de cet intérêt et intègre quelques fonctionnalités qui sont absentes d'autres systèmes de reconnaissance du domaine public, comme par exemple l'intégration de règles de prononciation, la possibilité d'utiliser une grammaire régulière arbitraire (et pas uniquement un modèle de langage classique), etc.

Après une présentation de l'état de l'art, nous présentons trois études qui portent sur la question de l'intégration des ressources linguistiques, dans le cadre successivement d'une application de reconnaissance de noms propres (section 4.2.2) ; puis du développement de deux systèmes de reconnaissance vocale grand vocabulaire (section 4.2.3 et section 4.2.4). Nous concluons en évoquant les travaux en cours et quelques perspectives.

4.2.1 Introduction : un rapide point sur l'état de l'art

Les systèmes de reconnaissance automatique de la parole sont passés, au cours de la dernière décennie, de l'état de prototypes de laboratoire à celui d'applications à destination du grand public, avec la commercialisation tout d'abord des systèmes de dictée vocale « grand vocabulaire », puis d'indexation de documents sonores radio ou télé-diffusés, avec le déploiement de systèmes de dialogue par téléphone utilisant la reconnaissance de parole, plus récemment avec l'apparition de systèmes de reconnaissance embarqués, dans des téléphones portables, des assistants personnels, ou encore des véhicules automobiles. De tels systèmes existent aujourd'hui pour la grande majorité des langues des pays industrialisés ou émergents et sont capables de fonctionner pratiquement en temps réel. La palette d'utilisation de ces systèmes s'est également considérablement élargie, accompagnant la constitution de grandes bases d'enregistrements audio et le besoin de disposer d'outils de recherche, d'extraction, de résumé, de traduction et de consultation appropriés.

[6] Mon activité en reconnaissance vocale doit beaucoup à une collaboration longue et soutenue avec G. Gravier (IRISA), avec qui j'ai collaboré à l'ENST, puis à IBM, et maintenant à distance. J'ai également une dette envers mes collègues du centre de recherche d'IBM à Yorktown Heights, en particulier envers M. Picheny, qui m'a invité à travaillé dans son groupe et envers G. Zweig, auprès duquel j'ai beaucoup appris.

Notons que ces nouvelles applications sont souvent construites par adaptation des outils de traitement des données textuelles aux sorties d'un système de reconnaissance vocale.

La technologie a donc évolué simultanément dans deux directions principales :
– prise en charge de conditions d'enregistrement de moins en moins contrôlées : reconnaissance de parole téléphonique, de parole dans un véhicule (voiture, avion...), d'interventions simultanées de plusieurs locuteurs...
– prise en charge de conditions et de styles d'élocution de moins en moins contraints : énoncés oralisés « spontanément », accents régionaux et ou étrangers, vocabulaires quasi illimités.

Plusieurs facteurs ont contribué à cette évolution : en plus de l'augmentation de la vitesse des processeurs et de l'espace mémoire disponible pour exécuter les applications de reconnaissance, retenons principalement :
– l'impulsion donnée par les campagnes d'évaluation organisées aux États-Unis sous l'égide du DARPA, et, à une échelle moindre, en France (sous l'égide de l'AUF entre 1996 et 2000, plus récemment dans le cadre du programme Français "Technolangue"). Une retombée positive de ces diverses actions a été la constitution et la diffusion de bases de données acoustiques de plus en plus volumineuses directement utilisables pour développer des systèmes.
– l'adoption, par l'ensemble de la communauté, d'un paradigme dominant : la reconnaissance de parole fondée sur des modèles statistiques [Jelinek et al., 1975], sur l'amélioration duquel s'est progressivement concentré l'essentiel des développements. Cette communauté de compréhension du problème a également favorisé la segmentation de la communauté et la spécialisation des équipes de recherche : une fois le cadre commun défini et globalement partagé, il devenait possible de se concenter sur un aspect particulier du système, les améliorations obtenues devenant potentiellement partageables avec le plus grand nombre.

La très large majorité des systèmes de reconnaissance vocale s'appuie sur le modèle probabiliste suivant [Jelinek, 1997]. Après identification des segments de parole, le signal de parole est converti en une suite d'observations acoustiques $x_1 \dots x_T$ par extraction d'une série de paramètres pertinents décrivant les caractéristiques locales du signal. Chaque observation acoustique est un vecteur dans $\mathbb{R}^n$, où n est typiquement de l'ordre de quelques dizaines. Si V dénote le vocabulaire et V^* l'ensemble de toutes les séquences de mots, la reconnaissance probabiliste s'exprime comme la recherche dans V^* de la séquence de mots $w_1^* \dots w_n^*$ la plus probable, conditionnellement à ce signal, donnant lieu aux programmes d'optimisation décrits dans les équations (4.3) et (4.4).

$$w_1^* \dots w_n^* = \underset{V^*}{\operatorname{argmax}}\, P(w_1 \dots w_n | x_1 \dots x_T) \tag{4.3}$$

$$= \underset{V^*}{\operatorname{argmax}}\, P(x_1 \dots x_T | w_1 \dots w_n) P(w_1 \dots w_n) \tag{4.4}$$

Le passage de (4.3) à (4.4) s'opère par application de la formule de Bayes, puisque le terme correspondant à la probabilité *a priori* du signal est le même pour toutes les hypothèses $w_1 \dots w_n$. La réalisation de ce programme demande la disponibilité de deux types de modèles statistiques :
– d'une modélisation de la probabilité d'une suite d'observations acoustiques correspondant à une séquence de mots donnée, fournie par les modèles acoustiques, conjoints à des dictionnaires de prononciation ;
– de la modélisation de la probabilité *a priori* d'une séquence de mots, fournie par le modèle linguistique.

Ces différents modèles constituent les ressources nécessaires au fonctionnement d'un décodeur de parole.

Des ressources et leur intégration

La réalisation du programme défini par l'équation (4.4) mobilise donc trois types de ressources :
– des dictionnaires de prononciation, associant aux mots du vocabulaire de reconnaissance une ou plusieurs représentation(s) phonétique(s) de leur prononciations ;

– des modèles acoustiques, reposant, dans la majorité des systèmes, sur l'utilisation de Modèles de Markov cachés (HMM) à lois d'émission multi-gaussiennes, chaque unité de représentation phonémique étant associée à un ou plusieurs HMM(s) particulier(s). Une forme usuelle est que chaque HMM corresponde à la modélisation statistique des réalisations d'un phone dans des contextes phonétiques gauches et droits donnés.
– des modèles linguistiques, pour l'essentiel construits par apprentissage statistique sur de larges corpus textuels, les modèles k-grammes (voir la section A.0.8) étant de loin les modèles les plus usités.

La combinaison de ces trois ressources permet de spécifier l'espace de recherche dans lequel la meilleure hypothèse de reconnaissance sera recherchée. Deux facteurs jouent alors un rôle crucial pour obtenir des systèmes performants et rapides :

– la bonne adéquation des ressources avec la tâche de décodage à effectuer ;
– la bonne intégration des ressources entre elles, permettant de développer des stratégies de recherche rapides et efficaces.

La première de ces préoccupations a fait l'objet d'une attention toute particulière, donnant lieu à une vaste littérature qu'il est vain de vouloir résumer ici, sur la problématique de *l'adaptation des ressources* (voir par exemple [Wellekens and Junqua, 2001, Junqua and van Noord, 2001] pour un échantillon représentatif des travaux conduits dans cette direction). Qu'il s'agisse d'adapter automatiquement les modèles acoustiques aux conditions d'enregistrement ou aux locuteurs, d'inférer automatiquement de nouvelles prononciations, ou de nouveaux mots à reconnaître, ou de régler les modèles linguistiques pour mieux prendre en compte les distributions lexicales liées à un thème particulier, la problématique est essentiellement la même. Dans tous les cas, l'adaptation vise à exploiter au mieux l'ensemble des données disponibles, en particulier celles qui sont les plus similaires à celles que traite l'application, en particulier celles qui sont les plus récentes (y compris les portions de signal en cours de reconnaissance et qui ne sont pas étiquetées), de manière à construire les modèles les plus pertinents possibles. Une mise en œuvre exemplaire de cette démarche consiste à démarrer avec des modèles généraux (ie. dont les paramètres sont appris sur de larges corpus plus ou moins indifférenciés), puis à transformer à la marge les (paramètres des) modèles utilisés par le décodeur à l'aide des nouvelles données disponibles pour l'adaptation.

La problématique de l'intégration des ressources est toute autre, et, d'une certaine manière, contradictoire avec la précédente : il s'agit ici essentiellement de proposer des méthodes de construction de l'espace de recherche sur lesquels pourront s'appuyer des algorithmes rapides et efficaces. Deux approches se sont progressivement imposées pour leur efficacité, qui partagent la même stratégie de recherche, à savoir l'algorithme de Viterbi [Viterbi, 1967], complété par des heuristiques d'élagage [Aubert, 2000]. Ces approches, sur lesquels nous aurons l'occasion de revenir plus en détail, sont les suivantes :

– dans [Ney and Aubert, 1996, Ney and Ortmanns, 1999], chaque prononciation (une suite d'unité phonétique) est remplacée par la suite des états des HMMs acoustiques correspondants. Cet ensemble (fini) de séquences d'états est représenté par son arbre des préfixes. L'espace de recherche est alors déployé dynamiquement au cours de la recherche de la manière suivante. L'arbre initial, contenant tous les mots observables en début de phrase, est progressivement traversé, au fur et à mesure du traitement des observations acoustiques. L'avancement de la recherche est matérialisé par un ensemble d'hypothèses, chaque hypothèse correspondant à un état actif de l'arbre et au passé linguistique h de cet état. Chaque nouvelle observation induit une transition dans le HMM et le développement des hypothèses les plus prometteuses à partir des hypothèses courantes. Lorsqu'une feuille de l'arbre est atteinte, marquant la fin de la reconnaissance du mot w, il est possible de cumuler au score courant la probabilité du modèle de langage $P(w|h)$, de mettre à jour l'histoire linguistique en y intégrant le dernier mot w et de développer de nouvelles hypothèses acoustiques en démarrant une nouvelle copie de l'arbre acoustique ; les hypothèses associées à cette copie partageant toutes le même passé linguistique hw (éventuellement tronqué).
  Cette approche se heurte toutefois à deux difficultés qui en compliquent l'implantation et en réduisent l'efficacité : les scores linguistiques ne sont appliqués que lorsqu'une fin de mot est atteinte ([Ortmanns et al., 1997] montre comment remédier à ce problème) ; l'utilisation de modèles acoustiques contextuels inter-

mot demande d'intégrer des mécanismes visant à limiter l'explosion combinatoire du nombre de nouvelles copies de l'arbre à considérer [Sixtus and Ney, 2002].

– [Pereira and Riley, 1996] pousse plus loin la logique de l'intégration, en s'appuyant sur le modèle des transducteurs finis. Constatant que, sous sa forme la plus usuelle, chacune des ressources est passible d'une modélisation par une machine à états finis, ces auteurs proposent essentiellement la construction suivante. Soient H le transducteur qui représente l'ensemble des modèles de Markov utilisés dans la modélisation acoustique, A le transducteur pondéré qui associe à chaque unité acoustique sa représentation phonétique, D le transducteur pondéré qui représente le dictionnaire de reconnaissance, et L l'automate fini pondéré représentant le modèle de langage [Riccardi et al., 1996], alors l'espace de recherche peut être construit en composant ces transducteurs selon : $H \circ A \circ D \circ L$. Il reste possible, dans ce cadre, d'utiliser les algorithmes traditionnels de recherche en faisceau (*beam search*). L'utilisation de techniques de déterminisation et de minimisation de ce transducteur pondéré [Mohri, 1997] permet d'optimiser cette construction et d'améliorer sensiblement la vitesse et la qualité du décodage [Mohri and Riley, 1998, Mohri et al., 1998].

  Les modèles à états finis sont également bien adaptés à la manipulation de graphes (ou treillis) de mots et ont favorisé l'adoption de stratégies d'apprentissage ou de décodage en plusieurs passes : chaque passe aboutit à une réduction de l'espace de recherche et permet l'utilisation de modèles plus fins et computationnellement plus coûteux.

  Cette approche a permis de construire des systèmes de reconnaissance fonctionnant en temps réel pour des très gros vocabulaires (plusieurs centaines de milliers de mots) et a montré qu'elle pouvait conduire à des performances meilleures que l'expansion dynamique de l'espace de recherche [Kanthak et al., 2002]. Elle pose toutefois un certain nombre de problèmes. De nombreux aménagements doivent être effectués pour s'assurer que les différentes étapes de la composition conduisent à des transducteurs compacts et déterminisables (par ajout de symboles auxiliaires, traitement particulier de certaines transitions spontannées...) [Mohri et al., 2002, Allauzen et al., 2004]. Le coût du calcul d'un transducteur compact pour $H \circ A \circ D \circ L$, qui implique la manipulation de machines composées de plusieurs millions, voire dizaines de millions d'états, limite la complexité des modèles acoustiques et linguistiques qui peuvent être intégrés [Chen, 2003]. Seconde faiblesse de ce type d'approche : le déchargement d'une large partie du coût computationnel du décodage sur des procédures effectuées hors-ligne annule les possibilités d'adaptation dynamique des ressources.

Les trois études présentées dans la suite de cette section s'inscrivent dans cette problématique d'intégration des ressources. Nous détaillons tout d'abord les grandes lignes d'un projet de reconnaissance de larges annuaires de noms propres, qui nous a conduit à réfléchir sur les stratégies de recherche dans un lexique de grande dimension ; nous évoquons ensuite le projet Sirocco, visant au développement d'un système ouvert de reconnaissance pour le grand vocabulaire, dont une des caractéristiques est la possibilité de prendre en compte des contraintes phonologiques durant le décodage ; nous détaillons enfin les travaux conduits au sein du groupe de recherche d'IBM, portant sur l'intégration de dépendances acoustiques à long terme dans un décodeur reposant sur la compilation de l'espace de recherche sous la forme d'un transducteur fini.

4.2.2 Reconnaissance de noms propres

Une des premières applications de la reconnaissance vocale pour la parole téléphonique a été l'automatisation de services d'annuaire ou de standard. Le scénario de ces applications est relativement simple : l'utilisateur prononce un nom, éventuellement complété par une localisation, et le système fournit le numéro ou bien encore met l'utilisateur en relation avec la personne recherchée.

En plus des difficultés intrinsèques de la reconnaissance vocale de parole téléphonique (étroitesse de la bande passante, variabilité des microphones et des conditions de bruit, etc.), une des difficultés de ce type d'application est la relative pauvreté du l'input, et, la taille des annuaires augmentant, la forte ambiguïté entre les noms répertoriés dans l'annuaire. Il est courant de devoir discriminer entre des noms qui ne diffèrent

que d'un seul phonème (e.g. *Beyssac-Peyssac*), sans parler des nombreux cas d'homophonie (e.g. *Dupont-Dupond*).

Le déploiement et l'acceptabilité de tels services demande donc d'avoir recours à des scénarios plus complexes, dans lesquels le système doit être, dans une certaine mesure, à même d'auto-évaluer la validitité de ses hypothèses [Jouvet and Monné, 1999], de manière à pouvoir demander à l'utilisateur, le cas échéant, de reformuler sa requête ou de fournir des informations complémentaires.

Le système sur lequel nous avons travaillé (voir [Neubert et al., 1998a,b] pour une description détaillée du système et une discussion des résultats obtenus) rentre dans ce cadre : en plus du nom de l'abonné, l'utilisateur doit également en fournir une épellation, de manière à augmenter les chances de retrouver le « bon » abonné. Le prototype développé repose sur une stratégie de décodage en plusieurs passes, s'appuyant sur une modélisation acoustique d'unités correspondant à des phones hors contexte (36 modèles) et à des lettres épelées (55 modèles[7])

– une première étape consiste en un décodage acoustique relativement « libre », la seule contrainte étant que le système doit d'abord reconnaître une série de phonèmes, puis un court silence optionnel, puis une série de lettres épellées. Chacun de ces types de séquences fait l'objet d'une modélisation unigramme. Cette première étape fournit le meilleur candidat acoustique, un candidat étant représenté par une prononciation et une épellation ;
– la seconde étape consiste à confronter ces candidats aux entrées de l'annuaire en calculant, pour chaque candidat, une distance d'édition avec toutes les entrées du lexique. Cette distance d'édition est une somme pondérée de la distance entre les parties correspondant à la prononciation d'une part, et des parties correspondant à l'épellation d'autre part. Les coûts d'insertion, de suppression et de substitution sont fondés sur des matrices de confusions acoustiques entre phonèmes. À l'issue de cette étape, l'entrée de l'annuaire qui se rapproche du meilleur candidat est finalement sélectionnée.

Ma contribution personnelle à ce projet a été double : j'ai d'une part constitué les ressources annuairiques nécessaires à l'apprentissage des modèles acoustiques et d'autre part travaillé sur l'étape de recherche lexicale. Les difficultés de cette tâche, ainsi que les procédés utilisés pour la réaliser sont longuement discutés dans [Yvon, 1996a, Béchet and Yvon, 2000] (voir également la section 4.1.1). Additionnellement, j'ai proposé d'augmenter très sensiblement (d'un facteur 10) la vitesse de l'étape de recherche lexicale, en la décomposant en deux temps : une étape de recherche approximative, qui, étendant l'approche originellement proposée par [Oflazer, 1996], trouve les n entrées annuairiques les plus proches au sens de la distance d'édition classique. Seuls ces n candidats sont alors comparés plus finement, selon les modalités décrites ci-dessus, avec l'hypothèse de sortie du décodeur de parole.

4.2.3 Autour de Sirocco

Le projet

Le projet Sirocco, initié en 1997, partait du besoin, partagé par de nombreux laboratoires français, de disposer d'un système de décodage de la parole continue grand vocabulaire sur lequel s'appuyer pour valider de nouveaux modèles. Un consortium, regroupant l'IRIT (Université de Toulouse), le LIA (Université d'Avignon), le LORIA (INRIA et Université de Nancy), l'IRISA (INRIA et Université de Rennes) et l'ENST s'est constitué, conduisant au développement de la première version de ce système, fondé sur les principes suivants :

– modularité du code, permettant un développement réparti sur plusieurs sites ;

[7] Et non 26, comme on s'y attendrait. La modélisation de stratégies d'épellation impose d'intégrer au répertoire des 26 lettres alphabétiques des mots comme : *deux*, utilisé pour épeler des lettres doubles ; *accent* ; *grave* ; *aigu*, utilisés pour les lettres accentuées, etc.

– ouverture et compatibilité avec les nombreux outils et formats communément adoptés dans la communauté : il ne s'agissait pas de redévelopper tous les outils nécessaires à la reconnaissance vocale, mais de se focaliser sur la tâche de décodage, en s'assurant de pouvoir accepter en entrée des ressources et de délivrer en sortie des objets manipulables par d'autres outils.

La stratégie de décodage implantée dans Sirocco est une recherche heuristique en faisceau (*beam search*) synchrone en temps, fondée sur une représentation du vocabulaire sous la forme d'un arbre des préfixes, suivant les travaux de [Ney and Aubert, 1996] (voir supra).

Ma contribution à ce projet, initialement focalisée sur les outils de manipulation des ressources lexicales (compilation du lexique et construction de l'arbre de décodage), s'est progressivement élargie au fur et à mesure du développement de la seconde version du système (intégration du modèle de langage, outils de manipulation des graphes de mots, coordination des développements logiciels), dans le cadre d'une collaboration avec l'IRISA de Rennes [Gravier et al., 2002]. La troisième version du système, développée en 2003/2004, a conduit à rendre le décodeur plus rapide et plus flexible, en le rendant à même de traiter des modèles de langage quelconques. Sa version actuelle, utilisée pour la campagne Ester d'évaluation des systèmes de reconnaissance et d'indexation d'émission radiodiffusées, permet le décodage en environ 5 fois le temps réel, pour des vocabulaires de reconnaissance de l'ordre de 65 000 mots.

D'autres évolutions sont actuellement en cours, visant à accéler le calcul des lois d'émissions des états acoustiques, à permettre l'utilisation de modèles acoustiques contextuels inter-mots, à intégrer des caractérisations phonétiques du signal, et à rendre possible un décodage rapide sur un graphe statique arbitraire (voir supra).

Le logiciel aujourd'hui existant correspond donc peu ou prou à l'état de l'art en matière de décodage acoustique, que ce soit en termes de rapidité de fonctionnement, d'espace mémoire requis ou de fonctionnalités disponibles. Il est compatible avec de nombreux autres outils du domaine public et est disponible sous licence GPL, librement téléchargeable, avec ses documentations, depuis la page du projet : `http://gforge.inria.fr/projects/sirocco`.

Décodage et intégration des ressources dans Sirocco

La qualité d'un système de reconnaissance de la parole dépend de la qualité des ressources utilisées lors du décodage. En ce qui concerne le lexique, les facteurs de qualité, en plus du taux de couverture, sont principalement liés à la qualité de la description des *variantes de prononciation*. L'influence de la qualité de description de la variation est en fait double, puisque ces variantes jouent également un rôle important lors de l'étape d'estimation des modèles acoustiques. Ces effets sont d'autant plus marqués que les énoncés à reconnaître se rapprochent de conditions d'élocution « spontanées » [Strik and Cucchiarini, 1999]. L'existence de variantes de prononciation, libres ou systématiquement conditionnées par des facteurs multiples (prosodiques, phonologiques, lexicaux...) est documentée, par exemple, dans [Lacheret-Dujour, 1990, Laporte, 1989]. La prise en compte de variantes de prononciation dans un système de reconnaissance ne peut, pour autant, se limiter à l'accumulation de nouvelles variantes, dans la mesure où l'ajout de variantes dans le lexique, si elle permet d'améliorer la reconnaissance des mots correspondants, est susceptible d'augmenter la confusion acoustique et par là de dégrader la reconnaissance d'autres mots [Hain, 2002] ; par ailleurs, l'utilité des variantes dépend de la précision des modèles utilisés par le décodeur (voir, par exemple, [Adda-Decker and Lamel, 1998, Jurafsky et al., 2001]).

Diverses propositions ont été faites pour (i) repérer les variantes de prononciation potentiellement intéressantes et pour (ii) les intégrer efficacement dans le lexique du décodeur. En ce qui concerne particulièrement ce dernier point, les stratégies suivantes ont été proposées dans la littérature :

– pondérer chaque prononciation par une probabilité de réalisation [Wester et al., 1998] ; une version plus ambitieuse consistant à estimer des modèles de langage portant non plus sur des formes graphiques,

mais sur des représentations phonétiques (eg. dans [Cremelie and Martens, 1995, Schiel et al., 1998]). Ces deux approches se heurtent toutefois au problème d'estimation de ces probabilités, en particulier pour les mots rares ;
– modéliser, dans le lexique, non plus des mots mais des multi-mots (*multi-word units*) : cette approche est adaptée à des séquences fréquentes, contenant des mots fonctionnels qui donnent souvent lieu à des effets de co-articulation importants [Riley et al., 1999, Kessens et al., 1999]. La question de la détection de ces unités reste toutefois délicate et la prise en compte de ces multi-mots a un effet potentiellement néfaste sur l'estimation des modèles de langage ;
– lorsque la variation est systématique, il est possible de la modéliser par des règles de variation, qui peuvent provenir d'une connaissance d'expert [Wester et al., 1998, Pérennou and Calmès, 2000], ou bien être inférées à partir des données [Ravishankar and Eskenazi, 1997, Riley et al., 1999, Wester and Fosler-Lussier, 2000]. Ces règles trouvent deux types d'utilisation : elles peuvent d'une part servir à générer des variantes non attestées en corpus, soit s'intégrer directement durant le décodage. En particulier, dans le cas où ces règles reposent sur des récritures contextuelles ordonnées appliquant sur des représentations phonétiques, il est possible de les représenter sous la forme de transducteurs à états finis [Kaplan and Kay, 1994] et de les intégrer dans un moteur de décodage à base de transducteurs à états finis [Smaili et al., 2002].

Le formalisme de représentation lexicale de Sirocco permet, d'une certaine manière, l'utilisation conjointe de ces trois stratégies, puisqu'il est possible d'associer, à chaque entrée lexicale, représentée comme une paire (w, p) associant une forme et une prononciation, (i) le contexte C (gauche et droit) dans lequel cette association est autorisée et (ii) un score exprimant la log-probabilité de cette association, correspondant à $\log(P(w|p, C))$. La description des contextes fait intervenir des classes lexicales de formes, et dans la dernière version, de classes de prononciations (voir la table 4.1).

(*)	*les*	(Vinit & Plural)	→	/lez/	0.0
(*)	*les*	(!(Vinit & Plural))	→	/le/	0.0

TAB. 4.1 – Exemple d'entrée lexicale intégrant des contraintes contextuelles : la forme *les*. L'opérateur ' !' dénote la négation.

La première de ces entrées exprime que la forme *les* peut se prononcer /lez/ dans un contexte gauche arbitraire ($*$) et suivie d'une forme appartenant simultanément aux classes lexicales $Vinit$ (présence d'une voyelle initiale) et $Plural$ (pluriel). Dans ces deux exemples, le score associé est arbitrairement fixé à 0.

L'utilisation de ces informations de contexte est prise en charge dynamiquement pendant le décodage : au terme du décodage de la prononciation p_i, associée au mot w_i, les contrôles suivants sont effectués :
– (i) (p_i, w_i) est un contexte droit compatible avec le choix réalisé pour la paire précédente (p_{i-1}, w_{i-1}) sur ce chemin d'hypothèses
– (ii) (p_{i-1}, w_{i-1}) est un contexte gauche acceptable pour la réalisation p_i du mot w_i.

Si ces deux conditions sont vérifiées, alors l'hypothèse courante pourra être développée, après intégration du score associée à cette variante de prononciation et du score linguistique ; dans le cas contraire elle sera supprimée. L'effet espéré de l'intégration de ces contraintes est donc de diminuer l'espace de recherche, en supprimant des chemins d'hypothèses invalides et par là d'améliorer les performances du décodeur.

[Gravier et al., 2001, 2004] présente en détail cette architecture ainsi que les résultats d'une série d'expériences conduites avec différents lexiques dérivés du lexique Mhatlex [Pérennou and Calmès, 2000], correspondant à une utilisation différentielle de ces contraintes de prononciation. Trois phénomènes ont été étudiés : l'élision des e-muets, la chute des consonnes liquides finales, les liaisons. Le résultat principal de ces études est que la prise en compte des contraintes de prononciation pendant le décodage est possible, mais qu'il induit un coût algorithmique évident, alors que l'amélioration des performances est loin d'être aussi nette. En plus de l'insuffisante qualité des modèles acoustiques et linguistiques utilisés pour ces expériences, cet échec relatif semble être dû principalement aux raisons suivantes :

– la description des contraintes est trop imprécise, en particulier elle n'utilise pas les informations lexicales liées à la partie du discours, qui sont pourtant importantes en particulier pour modéliser correctement les contextes de liaison ;
– plus fondamentalement, les contraintes de prononciation sont, dans nos expériences, imparfaitement compatibles avec le décodage : à l'instant t, des différentes hypothèses de fin de mot correspondant à la reconnaissance du mot w, et associées aux prononciations $p_1 \dots p_k$, une seule (celle qui a le meilleur score) est développée et donne lieu à une nouvelle copie de l'arbre lexical. Or, il faudrait attendre *la fin de la reconnaissance du mot suivant*, et donc l'expansion du contexte droit, pour décider quelles sont les prononciations qui peuvent effectivement être développées le long d'un chemin. Un calcul plus correct imposerait, en fait, de développer une nouvelle copie d'arbre pour chaque prononciation, donnant lieu à un accroissement important du nombre d'hypothèses à conserver.

Dans les dernières versions de Sirocco, il a donc été décidé de ne plus prendre en compte les contraintes de prononciation durant la première passe de décodage, mais de les utiliser dans une seconde passe, lors de la réévaluation des graphes de mots.

4.2.4 Une stratégie d'optimisation du graphe de décodage

Contexte

L'alternative principale à l'algorithme de décodage implanté dans le système Sirocco consiste à utiliser, dans des conditions de reconnaissance de grands vocabulaires, les mêmes stratégies que celles utilisées pour la reconnaissance de mots isolés. Comme expliqué à la section 4.2.1, l'espace de recherche, correspondant au graphe complet des états acoustiques accessibles, est alors intégralement précompilé hors-ligne. Le décodage consiste simplement à utiliser l'algorithme de Viterbi, complété par des heuristiques visant à limiter le nombre d'hypothèses (recherche en faisceau). La compilation hors-ligne de ce graphe mobilise essentiellement des outils de manipulation de transducteurs finis pondérés, les différentes ressources impliquées durant le décodage étant susceptibles d'être représentées sous cette forme.

Construction du réseau de décodage

L'adaptation[8] d'une stratégie de décodage utilisant des graphes intégralement pré-compilés à un système dans lequel les modèles acoustiques utilisent des contextes larges, comme c'est le cas dans les systèmes d'IBM, demande d'aménager la procédure de construction du graphe de décodage. Utiliser des contextes larges impose en particulier d'accorder une attention particulière à la construction du transducteur A, qui associe des noms d'unités acoustiques aux unités phonétiques correspondantes. Dans le cas le plus usuel où les unités acoustiques correspondent à des phones en contexte, ce transducteur a un état différent pour chaque paire de phones [Mohri et al., 2002]. Comme le note [Chen, 2003], l'utilisation de fenêtres de contexte plus larges conduit donc à une croissance exponentielle du nombre d'états de A.

Nous nous sommes particulièrement intéressés à la construction d'un réseau de décodage statique dans lequel les unités acoustiques sont calculées à partir de fenêtres de contexte intégrant le prédecesseur du mot courant. En d'autres termes, la séquence de modèles acoustiques $a(w)$ pour chaque mot w dépend de l'intégralité de la prononciation $l(v)$ du mot v à gauche de w.

Formellement, si l'on note $A(V)$ l'ensemble des représentations acoustiques de tous les mots du vocabulaire, et $L(V)$ l'ensemble de toutes les prononciations, cette dépendance est modélisée par un graphe biparti $G = (X, E)$ dont les sommets X sont bipartitionnés en respectivement $L(V)$ et $A(V)$, chaque arête $e =$

[8] Ces travaux ont été conduits au Watson Research Center d'IBM à Yorktown, New York, en étroite collaboration avec G. Zweig et G. Saon.

$(l(v), a(w))$ de E exprimant que la séquence $a(w)$ modélise le mot w dans le contexte gauche $l(v)$. Ce graphe joue le même rôle que le transducteur réalisant la composition $H \circ A \circ D$ dans l'approche décrite précédemment à la section 4.2.1 : chaque transition est étiquetée en entrée par une distribution de probabilité sur les observations acoustiques et les sorties correspondent aux mots du vocabulaire. Pour des applications typiques de reconnaissance vocale grand vocabulaire, la taille de ce graphe est imposante, puisque chaque sommet de $L(V)$ possède exactement $|L(V)|$ arcs sortants, conséquence du fait que le modèle de langage accepte chaque succession de deux mots quelconques du vocabulaire. G possède au total de l'ordre de $|V|^2$ arêtes, ce qui interdit même que G puisse être intégralement construit et stocké sur disque.

G peut toutefois être singulièrement compacté en factorisant les contextes gauches qui imposent des contraintes similaires et qui vont partager les mêmes successeurs. Cette factorisation se fonde sur l'ajout de nouveaux sommets dans le graphe, grâce auxquels il est possible de réduire considérablement le nombre d'arêtes dans le graphe (voir la figure 4.1).

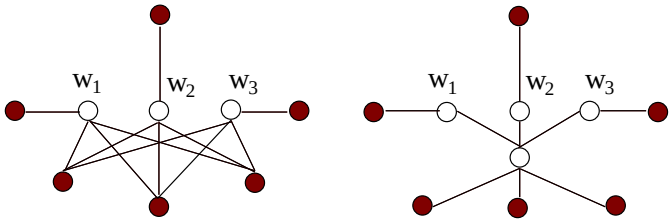

FIG. 4.1 – Le graphe des prononciations avant et après factorisation

La factorisation schématisée sur la figure 4.1 correspond formellement à la construction de bi-cliques (graphes bipartis complets) couvrant les arêtes de G. Si l'on note $B_i = (L_i, A_j)$ une bi-clique, et $n(B_i)$ son nombre de sommets, alors la recherche d'une factorisation optimale, contenant un nombre minimal d'arêtes, s'exprime par la recherche d'un ensemble de k bi-cliques telles que :

$$\forall e \in E, \exists i \text{ tq. } e \in B_i \tag{4.5}$$

$$\sum_{i=1}^{i=k} n(B_i) \text{ est minimum} \tag{4.6}$$

ou, en d'autres termes, la recherche de la couverture d'ordre minimale de G par des bi-cliques.

En nous appuyant sur une preuve de [Feder and Motwani, 1991], adaptée d'un résultat dû à [Holyer, 1981], nous avons montré dans [Zweig et al., 2002, Yvon et al., 2004b], que ce problème était NP-difficile. Nous avons également proposé diverses heuristiques pour construire des factorisations acceptables à partir d'une présentation incrémentale de G et développé des outils pour intégrer les dépendances ainsi modélisées avec des modèles statistiques de langage. Les transducteurs ainsi construits, ont après minimisation et déterminisation, été utilisés pour des tâches de reconnaissance grand vocabulaire (utilisant des dictionnaires d'environ 20 000 mots). Par ces techniques, nous avons finalement pu mettre en œuvre des systèmes fonctionnant en dessous du temps réel et tout à fait compétitifs, en termes de performances, avec les systèmes à l'état de l'art pour des tâches de reconnaissance de conversations téléphoniques [Saon et al., 2003].

L'intégration de telles techniques de construction de l'espace de recherche et de décodage dans le système Sirocco présenté à la section précédente fait partie des perspectives de développement de ce système.

4.2.5 Perspectives : vers une meilleure utilisation des ressources

Au terme de ce rapide survol de l'état de l'art et de nos propres contributions, essayons de dresser quelques perspectives, dans un contexte général marqué par l'augmentation de la taille des ressources disponibles

et l'augmentation des capacités de calcul. Pour les aspects qui nous préoccuppent, à savoir les modèles de prononciation, les modèles linguistiques et le décodage, il nous semble qu'un problème encore aujourd'hui mal résolu est celui de l'intégration des ressources, problème que nous développons dans deux directions dans les paragraphes qui suivent : autour de la question de l'optimisation de l'espace de recherche, puis autour de la question des post-traitements linguistiques.

Optimisation de l'espace de recherche

Ressources La construction de l'espace de recherche du décodage mobilise trois types de ressources : des modèles acoustiques, des dictionnaires et des modèles linguistiques. Ces trois ressources sont, le plus souvent, apprises et optimisées séparément, sur des corpus d'apprentissage distincts, optimisations qui impliquent de nombreuses décisions locales, effectuées de manière indépendante : nombre et topologie des HMMs, sélection du vocabulaire, paramètres du modèle linguistique (lissage et filtrage des comptes...). Dans des situations réelles, chacune de ces ressources résulte elle-même d'une combinaison de modèles appris sur des corpus distincts : ainsi dans le cas de modèles linguistiques, on combinera des modèles généraux entraînés sur de larges corpus généralistes avec des modèles spécifiques entraînés sur des textes proches des énoncés qui doivent être reconnus.

Au final, le réseau de décodage inclut donc des ressources complexes et disparates, la contribution de chacune aux résultats finaux reste difficile, sinon impossible, à démêler avec précision. Pourtant, il apparaît clairement que, par exemple, la finesse de la modélisation des variantes de prononciation doit dépendre des modèles acoustiques utilisés : des modèles faiblement discriminants seront d'ailleurs gênés, plus qu'aidés, par une modélisation phonétique trop riche. Comme le mentionnent [Adda et al., 1998] :

> How many times were the new prononciation variants, which were added to solve a given acoustic modeling problem, globally ineffective ? While solving the problem for which they were designed, variants often introduce new errors elsewhere, cancelling the local benefit : as variants may increase homophones rates, they become potential error sources.
>
> (...) An increasing number of Context Dependant acoustic models tends to reduce the need for pronunciation variants

De même, les modèles linguistiques ont pour fonction première d'aider le décodage à opérer les bonnes décisions en cas de confusion acoustique [Printz and Olsen, 2002] : il est donc raisonnable de penser que des modèles acoustiques différents bénéficieront inégalement d'un surcroît de finesse de la description linguistique.

Paramètres La construction de l'espace de recherche implique également le réglage d'une multitude de paramètres, visant à compenser de diverses manières les biais induits par la construction de scores agrégeant des vraisemblances calculées sur des modèles probabilistes hétérogènes :

– l'implantation pratique de l'équation (4.4) fait intervenir un facteur de pénalisation du score linguistique, de manière que sa plage de variation soit du même ordre de grandeur que les scores (vraisemblances) acoustiques ; la même correction devant s'appliquer aux probabilités des variantes de prononciation ;
– la combinaison d'un score linguistique et d'un score acoustique ne s'applique que lors d'une transition entre deux mots, ce qui conduit à donner une prime, pendant la recherche, aux hypothèses qui ne considèrent que peu de transitions inter-mots. Pour compenser cet effet, un facteur de pénalisation doit être spécifié, qui permet de réguler le débit (en mots) du décodeur ;
– enfin, la modélisation des pauses (courtes) et silences (longs) pendant le décodage est également cruciale : pourtant, les modèles linguistiques estimés sur des corpus textuels sont dans l'incapacité de prédire correctement les insertions de ces « mots » particuliers et les paramètres correspondants doivent faire l'objet d'un ajustement manuel.

Une proposition Au final, le développement d'un système dans un cadre applicatif donné mobilise donc des multitudes de petits ajustements, visant à optimiser séparément ou conjointement de multiples ressources et paramètres, chacun de ces choix pouvant avoir une influence très sensible sur les performances du décodage. Cette étape de réglage des paramètres donne lieu à un travail empirique ingrat, consistant à calibrer des valeurs des paramètres en observant les performances du système de reconnaissance sur des corpus de développement.

L'approche que nous préconisons vise deux objectifs : (ii) améliorer l'intégration des ressources et (ii) fournir des critères objectifs pour cette étape de calibrage. Le travail de thèse de Lin Shiuan-Sung, dont des résultats préliminaires sont donnés dans [Shiuan-Sung and Yvon, 2005], s'inscrit dans cette problématique. L'idée directrice de ce travail, dans la lignée des travaux de [Kuo et al., 2002, Roark et al., 2004], est de mettre en œuvre des techniques d'apprentissage discriminant (cf. la discusion de la section 1.3.3), dans notre cas des techniques de minimisation du taux d'erreur (MCE) [Juang et al., 1997] visant à optimiser directement les paramètres du réseau de décodage. Nous soutenons en effet que, du fait de la disponibilité de techniques de décodage très rapides (infra temps-réel), de telles stratégies d'estimation sont devenues aujourd'hui envisageables, à un coût computationnel qui, quoiqu'important, n'excède pas celui de l'étape d'ajustement empiriques des divers paramètres et ressources du décodeur.

Cette stratégie présente deux avantages :
– l'optimisation des ressources et des paramètres du système est opérée de manière conjointe et non plus indépendante : ainsi, par exemple, les différents scores linguistiques sont réglés en tenant compte de la capacité de discrimination des modèles acoustiques. Un seul critère d'optiminisation est utilisé et explicitement minimisé : le taux d'erreur du système de reconnaissance.
– il devient possible de construire des réseaux de décodage qui sont hors de portée des techniques de recombinaison classique. Prenons l'exemple des variantes de prononciation : dans une approche classique, chaque variante de prononciation d'un mot est nantie d'une probabilité hors-contexte, probabilité qui est utilisée pour pondérer le choix de chaque variante dans tous les contextes dans lesquels un mot peut s'insérer. À l'inverse, l'approche que nous préconisons permet, après apprentissage, de donner des scores différents à des variantes d'un même mot suivant les contextes dans lesquels il apparaît.

Décodage et post-traitements linguistiques

Avec l'amélioration et la diffusion des technologies de reconnaissance vocale, la gamme des contextes applicatifs s'est progressivement élargie : au côté des traditionnelles applications de commande vocale, de dictée et de dialogue, de nouveaux débouchés se sont installés qui recouvrent la gamme du champ d'application du traitement automatique des langues : indexation, recherche et extraction d'information, résumé, traduction. Ces applications visent à l'enrichissement ou à l'exploitation de corpus audio des plus variés : conversations et messages téléphoniques, émissions radio/télé, enregistrements de cours, de conférences, de procès, de réunions de travail, etc.

Paradoxalement, cette évolution n'a eu qu'un faible impact sur la conception des systèmes de reconnaissance. Suite notamment à la série de workshops organisés autour du corpus *Broadcast News*, s'est imposée l'idée qu'il suffirait, pour répondre à ces nouveaux besoins applicatifs, de mettre en place un *pipe-line*, la transcription automatiquement produite par un système classique servant d'entrée à des outils de traitement (robuste) des langues naturelles. Cette observation est sans doute justifiée pour les applications d'indexation, de routage et de filtrage, pour lesquelles les erreurs du décodeurs ne perturbent qu'à la marge la représentation en « sac-de-mots ». Il nous semble toutefois nécessaire de dépasser ce constat frustrant et qu'au moins deux pistes de réflexions mériteraient d'être poursuivies, notamment dans la perspective des nouvelles applications de la reconnaissance vocale (traduction, résumé, etc.).

Intégration précoce de ressources linguistiques La mise en place de chaînes de traitement impliquant des systèmes de reconnaissance et des outils de traitement automatique des langues pose la question de l'intégration des ressources mises en jeu par les différentes étapes de traitement. Par exemple, la détection d'entités nommées dans le flux de parole peut mobiliser, en plus du dictionnaire de reconnaissance, un dictionnaire listant les entitées connues, prenant le plus souvent la forme de mots composés. La question est alors de savoir s'il est possible d'utiliser efficacement ces mots composés dès l'étape de reconnaissance. Si, de surcroît, le détecteur d'entités est lui-même fondé sur un modèle statistique (exemplairement un HMM, voir la discussion de la section A.0.9), alors ne peut-on pas envisager d'intégrer ces modèles au fonctionnement du décodeur ? De même, si l'application visée consiste à détecter des thèmes, ou à suivre des évènements dans un flux audio, la question est d'utiliser au mieux les modèles de clustering ou de classification dès le décodage, en particulier pour améliorer la recherche et éviter de devoir utiliser des modèles linguistiques complexes et généraux pour reconnaître des énoncés qui sont très marqués thématiquement.

De nouvelles ressources « légères » De nombreux contextes applicatifs présentent la particularité qu'au côté des données audio, d'autres ressources sont potentiellement disponibles : dans le cas des conférences, un texte de présentation ou encore des transparents ; dans le cas des réunions un planning de la réunion, ou encore des documents qui seront discutés. Allant du plus proche (un support écrit de la conférence) au plus éloigné (des documents préparatoires d'une réunion), du plus pauvre linguistiquement (le programme d'une réunion) au plus riche (des supports de présentation), l'exploitation de ces écrits ouvre de nouvelles perspectives pour les systèmes de reconnaissance, autour notamment de la question de l'adaptation rapide et « légère » (au sens où les écrits disponibles ne sont pas toujours exploitables statistiquement) des systèmes de reconnaissance au thème des énoncés. Nous présentons le résultats de travaux préliminaires portant sur la reconnaissance de conférences dans le domaine juridique dans [Moissinac et al., 2004]. Ces travaux se poursuivent dans le cadre du réseau d'excellence K-Space.

Bibliographie

Gilles Adda, Josette Lecomte, Jospeh Mariani, Patrick Paroubek, and Martin Rajman. The GRACE French part-of-speech tagging evaluation task. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 1998.

Martine Adda-Decker and Lori Lamel. Pronunciation variation across systems, languages and speaking styles. In *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 1–6, Rolduc, The Netherlands, 1998.

Cyril Allauzen, Mehryar Mohri, Brian Roark, and Michael Riley. A generalized construction of integrated speech recognition transducers. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montréal, Canada, 2004.

Xavier Aubert. A brief overview of decoding techniques for large vocabulary continuous speech recognition. In *Proceedings of the ISCA Tutorial and Research Workshop, Automatic Speech Recognition : Challenges for the new Millenium (ASR2000). Paris, France, September 2000.*, pages 91–96, Paris, France, 2000.

Valérie Beaudouin and François Yvon. Contribution de la métrique à la stylométrie. In *Actes des 7èmes Journées Internationales d'Analyse Statistique des données textuelles (JADT)*, volume 1, pages 107–118, Louvain la Neuve, Belgique, 2004.

Valérie Beaudouin and François Yvon. The metrometer : a tool for analysing the French verse. *Literary and Linguistic Computing*, 11(1), 1996.

Frédéric Béchet. LIA_PHON : Un système complet de phonétisation de textes. *Traitement Automatique des Langues*, 42(1) :47–67, 2001.

Frédéric Béchet and François Yvon. Les noms propres en traitement automatique de la parole. *Traitement Automatique des Langues*, 41(3) :671–707, 2000.

Alan W. Black and Paul Taylor. CHATR : a generic speech synthesis system. In *Proceedings of the International Conference on Computational Linguisitics (COLING)*, volume 2, pages 983–986, Kyoto, Japan, 1994.

Alan W. Black and Paul A. Taylor. The Festival Speech Synthesis System : System documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, 1997. Avaliable at http ://www.cstr.ed.ac.uk/projects/festival.html.

Philippe Boula de Mareuïl, Véronique Aubergé, Jacqueline Vaissière, and François Yvon. *Guide du transcripteur pour la constitution du lexique phonétique MultPhon*. Phase 1 du programme “Ingénierie des Langues”(thème 3) du CNRS, 1998.

Philippe Boula de Mareüil, Christophe d’Alessandro, François Yvon, Jacqueline Vaissière Véronique Aubergé, and Angélique Amelot. A french phonetic lexicon with variants for speech and language processing. In *Proceedings of the 2nd Language Resources Engineering Conference (LREC)*, volume I, pages 273–276, Athens, Greece, 2000.

Nick Campbell and Alan W. Black. Prosody and the selection of source units for concatenative synthesis. In Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors, *Progess in Speech Synthesis*, pages 279–292, New York, 1996. Springer Verlag.

Nina Catach. *La phonétisation automatique du Français*. Éditions du CNRS, Paris, 1984.

Stanley F. Chen. Compiling large context phonetic decision trees into finite-state transducers. In *Proceedings of Eurospeech/Interspeech 2003*, volume 2, pages 1169–1172, Geneva, Switzerland, 2003.

Patrick Constant. *Analyse syntaxique par couche*. PhD thesis, École Nationale Supérieure des Télécommunications (ENST), Paris, France, 1991.

Nick Cremelie and Jean-Pierre Martens. On the use of pronunciation rules for improved word recognition. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 3, pages 1747–1750, Madrid, Spain, 1995.

Philippe Boula de Mareüil. *Étude linguistique appliqueée à la synthèse de la parole à partir du texte*. PhD thesis, Université Paris XI, Orsay, 1997.

Sabine Deligne, François Yvon, and Frédéric Bimbot. Introducing statistical dependencies and structural constraints in variable-length sequence models. In Laurent Miclet and Colin de la Higuera, editors, *Grammatical Inference : Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence 1147, pages 156–167. Springer, September 1996.

Michel Divay. A written text processing expert system for text to phoneme conversion. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, pages 853–856, Kobe, Japan, 1990.

Robert E. Donovan. *Trainable Speech Synthesis*. PhD thesis, University of Cambridge, 1996.

Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, NL, 1997.

Tomás Feder and Rajeev Motwani. Clique partitions, graph compression and speeding-up algorithms. In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, New Orleans, Louisiana, 1991*, pages 123–133, 1991.

Pierre Fouché. *Traité de prononciation française*. Klincksieck, Paris, 1959.

Andrew R. Golding. *Pronouncing names by a combination of rule-based and case-based reasoning*. PhD thesis, Stanford University, Stanford, CA, 1991.

Guillaume Gravier, François Yvon, Bruno Jacob, and Frédéric Bimbot. Integrating contextual phonological rules in a large vocabulary decoder. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 4, pages 2293–2296, Aalborg, Danemark, 2001.

Guillaume Gravier, François Yvon, Bruno Jacob, and Frédéric Bimbot. Sirocco, un système ouvert de reconnaissance de la parole. In *Actes des XXIVe Journées d'Etudes sur la Parole (JEP'02)*, Nancy, France, 2002.

Guillaume Gravier, François Yvon, Bruno Jacob, and Frédéric Bimbot. Integrating contextual phonological rules in a large vocabulary decoder. In Wim van Dommelen and William Barry, editors, *Integration of Phonetic Knowledge In Speech Technology*. Kluwer Academic Press, 2004.

Thomas Hain. Implicit pronounciation variation in ASR. In *Proceedings of ITRW PMLA 2002*, Estes Park, Colorado, 2002.

Ian Holyer. The NP-completeness of some edge partition problems. *Siam Journal of Computer Science*, 10 (4) :713–717, 1981.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, MA, 1997.

Frederik Jelinek, Lalit R. Bahl, and Robert L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21 :250–256, 1975.

Denis Jouvet and Jean Monné. Recognition of spelled names over the telephone and rejection of data out of the spelling lexicon. In *EUROSP*, pages 283–286, Budapest, Hungary, 1999.

Biing-Hwang Juang, Wu Chou, and Chin-Hui Lee. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio processing*, 5(3) :266–277, 1997.

Jean-Claude Junqua and Gertjan van Noord, editors. *Robustness in Language and Speech Technology*, volume 17 of *Text, Speech and Language Technology*. Kluwer Academic Publishers, 2001.

Daniel Jurafsky, W. Ward, J. Zhang, K. Herold, X. Yu, and S. Zhang. What kind of pronunciation variation is hard for triphones to model ? In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 577–580, Salt Lake City, UT, 2001.

Stephan Kanthak, Hermann Ney, Michael Riley, and Mehryar Mohri. A comparison of two LVR search optimization techniques. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, pages 1309–1312, Denver, CO, 2002.

Ronald M. Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3) :331–378, 1994. (First appeared as a paper presented to the Winter Meeting of the Linguistic Society of America, New York, 1981.).

Judith M. Kessens, Mirjam Wester, and Helmer Strik. Modeling within-word and cross-word pronunciation variation to improve the performance of a Dutch CSR. In *Proceedings of the International Congress of Phonetic Science (ICPhS)*, pages 1665–1668, San Francisco, CA, 1999.

Hong-Kwang Jeff Kuo, Eric Fosler-Lussier, Hui Jiang, and Chin-Hui Lee. Discriminative training of language models for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 325–329, Orlando, FL, 2002.

Anne Lacheret-Dujour. *Contribution à une analyse de la variabilité phonologique pour le traitement automatique de la parole continue multi-locuteur*. PhD thesis, Université de Paris VII, 1990.

Éric Laporte. Quelques variations phonétiques en français. *Linguisticæ Investigationes*, XII(1) :43–116, 1989.

Mehryar Mohri. Transducers in language and speech. *Computational Linguistics*, 23(2) :269–311, 1997.

Mehryar Mohri and Michael Riley. Network optimisation for large vocabulary speech recognition. *Speech Communication*, 25(3) :1–12, 1998.

Mehryar Mohri and Richard W. Sproat. An efficient compiler for weighted rewrite rules. In *Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL)*, pages 231–238, 1996.

Mehryar Mohri, Michael Riley, Don Hindle, Andrej Ljolje, and Fernando Pereira. Full expansion of context-dependent networks in large vocabulary speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 665–668, San Francisco, CA, 1998.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1) :69–88, 2002.

Jean-Claude Moissinac, François Yvon, and Slim Ben Hazez. Automating indexing of classes and conferences. In *the Proceedings of RIAO 2004*, Avignon, France, 2004.

Franck Neubert, Guillaume Gravier, François Yvon, and Gérard Chollet. Reconnaissance de noms prononcés et épelés au téléphone par modèles de markov cachés. In *Actes des XII journées d'études sur la parole (JEP)*, Martigny, Suisse, 1998a.

Franck Neubert, Guillaume Gravier, François Yvon, and Gérard Chollet. Directory name retrieval over the telephone in the picasso project. In *Proceedings of the IEEE IVTTA'98 workshop*, Torino, Italy, 1998b.

Herman Ney and Xavier Aubert. Dynamic programming search strategies : from digit strings to large vocabulary word graphs. In Chuin-Hui Lee, Franck K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, pages 385–411. Kluwer Academic Press, 1996.

Hermann Ney and Stefan Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, pages 64–83, 1999.

Kemal Oflazer. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1) :73–89, 1996.

Stefan Ortmanns, Andreas Eiden, Hermann Ney, and Norbert Coenen. Look-ahead techniques for fast beam search. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1783–1786, 1997.

Fernando C. N. Pereira and Michael D. Riley. Speech recognition by composition of weighted finite automata. In Emmanuel Roche and Yves Schabes, editors, *Finite-state natural language processing*. The MIT Press, 1996.

Guy Pérennou and Martine De Calmès. MHATLex : Lexical resources for modeling the French pronunciation. In *2nd International Conference on Language Resources and Evaluation (LREC)*, volume 1, pages 257–264, Athens, Greece, 2000.

Guy Pérennou, Daniel Cotto, Monique de Calmès, Isabelle Ferrané, Jean-Marie Pécatte, and J. Tihoni. Composantes phonologique et orthographique de BDLEX. In *Actes des 2èmes journées du GRECO-PRC CHM*, Toulouse, France, 1991. EC2 Editeur.

Marc Plénat. Vers une oralisation des sigles. In *Actes des IIèmes Journées internationales du GRECO-PRC Communication Homme-Machine*, pages 363–371. EC2 Eds, Nanterre, 1992.

Harry Printz and Peder Olsen. Theory and practice of acoustic confusability. *Computer, Speech and Language*, 16(1) :131–164, 2002.

Mosur Ravishankar and Maxime Eskenazi. Automatic generation of context-dependent pronunciations. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 5, pages 2467–2470, Rhodes, Greece, 1997.

Yann Régis-Gianas and François Yvon. Dimi ! In *Actes des XXVe journées d'études sur la parole (JEP)*, pages 421–425, Fez (Maroc), 2004.

Guiseppe Riccardi, Roberto Pieraccini, and Enrico Bocchieri. Stochastic automata for language modeling. *Computer, Speech and Language*, 10(265–293), 1996.

Michael Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John McDonough, Helen Nock, Murat Saraclar, Charles Wooters, and George Zavaliagkos. Stochastic pronunciation modeling from hand-labelled phonetic corpora. *Speech Communication*, 29(2–4) :209–224, 1999.

Brian Roark, Murat Saraclar, and Michael Collins. Corrective language modeling for large vocabulary asr with the perceptron algorithm. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 749–752, Montreal, Canada, 2004.

George Saon, Geoffrey Zweig, Brian Kingsbury, Lidia Mangu, and Upendra Chaudhari. An architecture for rapid decoding of large vocabulary conversational speech. In *Proceedings of Eurospeech/Interspeech 2003*, page to appear, Geneva, 2003.

Florian Schiel, Andreas Kipp, and H. G. Tillman. Statistical modeling of pronunciation : it's not the model, it's the data. In *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 131–136, Rolduc, The Netherlands, 1998.

Lin Shiuan-Sung and François Yvon. Discriminative training of finite-state decoding graphs. In *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, pages 733–736, Lisbon, Portugal, 2005.

Achim Sixtus and Hermann Ney. From within-word model search to across-word model search in large vocabulary continuous speech recognition. *Computer, Speech and Language*, 16 :245–271, 2002.

Nasser Smaili, Patrick Cardinal, Gilles Boulianne, and Pierre Dumouchel. Disambiguation of finite-state transducers. In *Proceedings of COLING'02*, Taipeh, 2002.

Richard W. Sproat, editor. *Multi-lingual Text-to-Speech Synthesis : the Bell labs apprach*. Kluwer Academic Publishers, Boston, MA, 1997.

Helmer Strik and Catia Cucchiarini. Modeling pronunciation variation for ASR : a survey of the literature. *Speech Communication*, 29(2–4) :225–246, 1999.

Tony Vitale. An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics*, 17(3) :257–276, 1991.

Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13 :260–269, 1967.

Christian Wellekens and Jean-Claude Junqua, editors. *ISCA Workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001.

Mirjam Wester and Eric Fosler-Lussier. A comparison of data-derived and knowlegde-based modeling of pronunciation variation. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, volume 1, pages 270–273, Beijing, China, 2000.

Mirjam Wester, Judith M. Kessens, and Helmer Strik. Modeling pronunciation variation for a Dutch CSR : Testing three methods. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, volume 6, pages 2535–2538, Sydney, Australia, 1998.

François Yvon. Règles de transcription graphème-phonème pour la prononciation automatique des sigles. *Lynx*, (30) :153–166, 1994.

François Yvon. *Prononcer par analogie : motivations, formalisations et évaluations*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 1996a.

François Yvon. Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks. In *Proceedings of the conference on New Methods in Natural Language Processing (NeMLaP II)*, pages 218–228, Ankara, Turkey, 1996b.

François Yvon, Philippe Boula de Mareüil, Christophe d'Alessandro, Véronique Aubergé, Michel Bagein, Gérard Bailly, Frédéric Béchet, Salia Foukia, Jean-Philippe Goldman, Eric Keller, Douglas O'Shaughnessy, Vincent Pagel, Frédérique Sannier, Jean Véronis, and Brigitte Zellner. Objective evaluation of grapheme-to-phoneme conversion for Text-To-Speech synthesis in French. *Computer, Speech and Language*, 12(4), 1998.

François Yvon, Philippe Boula de Mareüil, Christophe d'Alessandro, Véronique Aubergé, and Jacqueline Vaissière. Ressource standard pour le francais : un large lexique orthogaphiphe-phonétique. In *Actes des XXIIIe Journées d'Etudes sur la Parole (JEP)*, Aussois, France, 2000.

François Yvon, Nicolas Stroppa, Arnaud Delhay, and Laurent Miclet. Solving analogies on words. Technical Report D005, Ecole Nationale Supérieure des Télécommunications, Paris, France, 2004a.

François Yvon, Geoffrey Zweig, and George Saon. Arc minimization in finite-state decoding graphs with cross-word acoustic context. *Computer Speech and Language*, 18(4) :397–415, 2004b.

François Yvon. Lire les sigles : pour une approche à base de contraintes. In *Actes du séminaire Lexique et Communication Parlée du PRC CHM*, Toulouse, France, 1996c.

Geoffrey Zweig, François Yvon, and George Saon. Arc minimization in finite-state decoding graphs with cross-word acoustic context. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, volume 1, pages 389–392, Denver, CO, 2002.

Annexe A

Quelques tâches d'analyse linguistique

Cette annexe présente un certain nombre de tâches d'analyse linguistique pouvant être typiquement effectuée par un module de traitement autonome, en détaillant principalement les entrées-sorties de ces modules et en insistant sur la manière dont ces tâches peuvent plus ou moins facilement se prêter à la mise en œuvre de techniques d'apprentissage automatique.

Cette présentation aborde successivement les principales étapes de l'analyse linguistique d'un texte ou d'un énoncé : segmentation et pré-traitements, analyse lexicale, analyse syntaxique et analyse sémantique, modules qui correspondent à des tâches *intermédiaires* dans la terminologie du domaine (cf. la section 1.1.1). Comme il est usuel, notre présentation fait l'hypothèse que ces modules de traitement s'enchaînent sous la forme d'un *pipe-line*, chaque module fournissant une sortie unique qui est passée au module suivant. Cette hypothèse est sans aucun doute une idéalisation, dans la mesure où certaines ambiguïtés qui apparaissent à un niveau de traitement ne pourront être en réalité être levées qu'au terme des traitements ultérieurs. Les exemples de telles ambiguïtés abondent, ainsi (tiré de [Ide and Véronis, 1998]) :

(5) *l'étagère ploie sous les livres,*

pour lequel la désambiguïsation du genre (et du sens de) *livre* demande une forme d'analyse sémantique de la phrase.

A.0.6 Segmentation et prétraitements

La première étape du traitement automatique de textes consiste à repérer les unités linguistiques (mots, phrases, etc.) dans ce qui se présente initialement comme un flux de caractères. Ce repérage ne va pas de soi pour au moins deux raisons :

– les textes réels comportent de nombreuses unités qui ne sont pas, à proprement parler, linguistiques, mais relèvent soit de conventions typo-dispositionnelles (par exemple, '-' pour introduire un élément d'une liste, '_' pour introduire un tour de parole, etc.) ; soit de catégories extra-linguistiques (résultats sportifs, indicateurs boursiers, résultats d´élections, montants, numéros de téléphone, adresses physiques et électroniques, etc.) dont la présentation et le formatage utilisent des conventions (abbréviations, disposition) variables d'une source textuelle à une autre ;
– la définition des unités linguistiques pose problème : d'une part, les séparateurs typographiques ne sont pas toujours des marqueurs stables et fiables d'une frontière ; d'autre part, les unités apparaissent sous des formes variables (avec ou sans majuscule, suivant la position dans la phrase ; avec ou sans '-' pour certains composés morphologiques ; sous forme complète ou abrégée ou siglée pour les noms de firmes ou d'organisation, etc.).

Les modules de prétraitement se trouvent donc confrontés simultanément à deux problèmes : détecter des frontières dans un flux et assigner un type aux unités reconnues. L'approche traditionnelle consiste à dé-

velopper des modules de prétraitement plus ou moins *ad-hoc* à base d'expressions rationnelles (voir eg. [Sproat, 1997] pour une discussion de ces problèmes dans un contexte de synthèse vocale).

Développer un module de prétraitement est une tâche ingrate et fastidieuse, qui requiert de nombreuses connaissances des spécificités et conventions du domaine. Ceci semble la prédestiner à être un champ d'application idéal pour des techniques d'apprentissage, d'autant plus qu'il est relativement « naturel » de l'exprimer comme une catégorisation binaire séquentielle.

Pourtant, hormis notamment [Sproat et al., 2001], qui présente une étude complète des problèmes de normalisation ainsi qu'une palette d'outils (de segmentation et de classification) pour y faire face, ainsi quelques études sur la détection des fins de phrases e.g. [Rathnaparkhi, 1998], les travaux sur ce thème restent relativement rares, reflétant (i) le fait que cette tâche est souvent considérée comme résolue et (ii) son évaluation est difficile, un accord même minimal sur les segments à identifier étant extrêmement difficile à établir.

A.0.7 Analyse lexicale, ressources lexicales

La mise en œuvre de systèmes de traitement automatique des énoncés présuppose la disponibilité de ressources lexicales, qui permettent (i) d'identifier les formes graphiques et (ii) de leur associer un certain nombre de propriétés : une ou plusieurs étiquettes morpho-syntaxiques, une prononciation, une structure interne, des propriétés combinatoires (eg. des cadres de sous-catégorisation), une liste de sens possibles, des équivalents de traduction, etc.

Ces informations figurent, sinon intégralement, du moins partiellement, dans les dictionnaires monolingues ou bilingues « traditionnels ». La numérisation de dictionnaires a donc été, pendant longtemps, le principal moyen pour constituer des lexiques fiables. Le triple constat que (i) les informations fournies par les lexicographes n'étaient pas toujours suffisamment complètes ou cohérentes dans une optique de traitement automatique ; que (ii) les dictionnaires ne permettaient pas d'obtenir une couverture exhaustive des formes rencontrées dans les énoncés naturels et que (iii) la définition des entrées lexicales et de leurs propriétés étaient susceptibles de fortes variations lorsque l'on passait de l'étude de la langue générale à celle de langues de spécialité, ont toutefois conduit à envisager la constitution, la structuration et l'extension de ressources lexicales comme une tâche importante du traitement automatique des langues.

Dans ce cadre comme ailleurs, deux pistes ont été principalement envisagées : la constitution de bases de connaissances par des experts et les approches à base de données. Cette seconde catégorie se subdivise elle-même en deux types de travaux : des travaux à base de corpus, qui visent à inférer les propriétés des mots à partir de leurs contextes d'usage ; et des travaux à base de dictionnaires, qui visent à généraliser des dictionnaires existants à de nouvelles entrées. Un cas limite étant constitué par des études visant à redéfinir, exclusivement à partir de données d'usage, la notion d'entrée lexicale et de catégorie syntaxique. Ainsi, par exemple, [Finch and Chater, 1992] propose de redéfinir les catégories syntaxiques en constituant des classes homogènes de formes à partir de leur comportement distributionnel.

Dans le reste de cette section, je présente deux sous-tâches qui entrent dans ce cadre d'acquisition de connaissances lexicales et sur lesquelles j'ai personnellement travaillé : la conversion orthographique phonétique et l'analyse morphologique. Ce parti pris ne doit pas occulter le fait que de nombreuses études se sont attachées à acquérir d'autres éléments de la description lexicale (sens, cadres de sous-catégorisation...), à l'aide, souvent, de techniques voisines de celles présentées ci-dessous.

La conversion graphème-phonème

La conversion graphème-phonème est usuellement envisagée sous l'angle du calcul d'une séquence d'unités représentant la prononciation d'un mot (généralement des unités phonétiques, éventuellement enrichies de

v	o	-	y	a	g	e	r
v	w	a	j	a	ʒ	e	-

TAB. A.1 – Alignement bi-univoque de *voyager*

marques prosodiques), à partir d'une représentation orthographique d'un mot. Cette étape de calcul constitue l'une des phases de traitement dans les systèmes de synthèse vocale ; elle permet également de construire des dictionnaires et des données d'entraînement pour l'apprentissage de modèles acoustiques en reconnaissance automatique de la parole ; elle a enfin une utilité pour des applications de correction orthographique.

L'approche « traditionnelle » de ce problème consiste à écrire des règles de récriture exprimant la correspondance (contextuelle) entre séquences alphabétiques et séquences phonétiques ([Divay and Vitale, 1997] est un représentant relativement récent de cette approche). Ces règles sont ordonnées de la plus à la moins spécifique et sont appliquées à une chaîne d'entrée en respectant strictement cet ordonnancement. Etant donnée une entrée orthographique w, parcourue de la gauche vers la droite, on recherche la première règle qui s´applique : elle est déclenchée, conduisant à la récriture d'un préfixe v de w. Ce processus est itéré sur le nouvel input $v^{-1}w$, jusqu'à épuisement total de w.

L'application de techniques d'apprentissage à cette tâche (voir [Yvon, 1996] pour un état de l'art un peu daté, par exemple [Bisani and Ney, 2003, Chen and Maison, 2003] pour des travaux plus récents) visant plus souvent à tester une méthodologie d'apprentissage particulière plutôt qu'à concevoir le module de prononciation d'un système complet, les simplifications suivantes sont généralement opérées :

– les entrées (graphiques) et les sorties (phonétiques) sont alignées terme à terme, comme dans l'exemple suivant (voir la figure A.1), nécessitant d'introduire un symbole « vide » (figuré ci-dessous par '-') pour compenser les différences de longueur. [Yvon, 1996] présente diverses manières de calculer un tel alignement et discute les difficultés pratiques et théoriques que pose ce calcul.
– l'entrée du système est réduite à la seule forme orthographique, bien que celle-ci soit souvent insuffisante pour déterminer une prononciation. Il existe ainsi, en français, plusieurs centaines de formes graphiques qui correspondent à plus d'un lemme, ambiguïté qui s'accompagne d'une incertitude sur la prononciation, qui est le plus souvent levée par l'identification de la catégorie syntaxique du lemme ;
– les entrées sont présentées hors de leur contexte, alors qu'il existe des phénomènes contextuels susceptibles d'influer sur la prononciation d'un mot (par exemple, en français, la liaison) ;
– les sorties se limitent à une chaîne segmentale, alors que des informations de nature supra-segmentale sont également requises dans un contexte de synthèse vocale ;
– une sortie unique est construite pour chaque entrée, bien que, notamment pour les applications de reconnaissance vocale, le calcul de *variantes de prononciation* soit également nécessaire.

Autour de l'analyse morphologique

Le titre volontairement hésitant donné à ce paragraphe exprime la difficulté à donner une présentation homogène et cohérente des différents travaux conduits sous l'étiquette « d'apprentissage de la morphologie ».

Analyser des formes inconnues D'un point de vue applicatif, l'analyse morphologique est principalement envisagée comme un outil de traitement des formes inconnues (entendre : absentes du vocabulaire de l'application) [Sproat, 1992]. Suivant les langues et les applications visées, le rôle de ce composant est d'autant plus important que le lexique original est petit et que la couverture des règles est grande. En général, l'entrée de l'analyseur se réduit à une séquence orthographique ; la sortie attendue est variable suivant les applications mais comprend au moins une décomposition en unités morphématiques, de laquelle il est possible de dériver le lemme (ou la base) de la forme inconnue et certaines caractéristiques morpho-

syntaxiques, dont, en particulier, la (ou les) catégories syntaxiques du lemme. Comme le décrit [Sproat, 1992], cette analyse requiert un grand nombre de connaissances linguistiques, en particulier une description de la *morphotactique* (des arrangements licites de morphèmes), ainsi que des règles morpho-phonologique (ou morpho-orthographiques) exprimant les modifications formelles découlant de la combinaison de deux morphèmes. PC-Kimmo [Koskenniemi, 1983, Antworth, 1990], référence dans le domaine de la morphologie computationnelle, dispose ainsi de deux mécanismes (un automate décrivant les suites de morphèmes licites et une grammaire d'unification opérant sur ces suites de morphèmes) pour exprimer les régularités morphotactiques ; ainsi que du mécanisme des règles à deux niveaux pour exprimer les changements formels. [Zajac, 2001] présente une stratégie d'apprentissage supervisé d'un système de règles morphotactiques et morpho-phonologiques complet à partir d'exemples étiquetés et sous le contrôle d'un expert humain. L'apprentissage de la morphologie est toutefois le plus souvent envisagé dans le cadre de sous-tâches moins ambitieuses.

Devins morpho-syntaxiques Une première sous-tâche relativement bien définie consiste à deviner la ou les catégories morpho-syntaxiques possibles pour une forme orthographique inconnue. De tels outils trouvent une application directe dans les étiqueteurs morpho-syntaxiques qui attendent en général qu'une telle information soit disponible, y compris pour les mots hors-vocabulaire. L'apprentissage de « devins » morpho-syntaxiques est discutée par exemple dans [Brill, 1994, Mikheev, 1997, Béchet et al., 1997] : pour l'essentiel la divination est modélisée comme une tâche de catégorisation, s'appuyant sur la base de caractéristiques typographiques, ainsi que par détection d'affixes caractéristiques ; elle intègre également, dans certains modèles, la prise en compte du contexte syntaxique d'occurrence.

Relations morphologiques Une seconde sous-tâche consiste à inférer des relations morphologiques flexionnelles ou dérivationnelles entre mots ou termes, de façon à pouvoir regrouper automatiquement sous une même entrée d'index des variantes morphologiques d'un même mot ou terme. En ce qui concerne la morphologie dérivationnelle, une telle projection est traditionnellement opérée dans les applications d'indexation et de recherche documentaire par le biais d'heuristiques tronquant suffixes et préfixes trouvés dans une liste pré-établie : on parle alors de *racinisation* (*stemming*). Cette approche a été popularisée dans la communauté de la recherche d'information par la diffusion d'algorithmes tels que celui de Porter [1980]. Des tentatives visant à induire automatiquement des raciniseurs à partir de lexiques ou de corpus sont présentées par exemple dans [Krovetz, 1993, Jacquemin, 1997]. [Gaussier, 1999] se situe dans une problématique identique ; la stratégie d'apprentissage qu'il met en œuvre va toutefois plus loin puisqu'après avoir induit, à partir d'un lexique et de manière non supervisée, une approximation des familles dérivationnelles, il propose une méthode pour également apprendre des règles de dérivation probabilisées en exploitant des co-occurrences entre des transformations graphiques et des changements de catégorie syntaxique. Des travaux voisins sont présentés dans [Schone and Jurafsky, 2000, Baroni et al., 2002, Schone and Jurafsky, 2001], qui infèrent des classes de formes morphologiquement apparentées, en se fondant non seulement sur des proximités orthographiques, mais également sur une proximité sémantique calculée à partir des contextes d'occurrence. [Goldsmith, 2001] poursuit des buts voisins, cherchant à construire simultanément et de manière non supervisée des classes de stemmes et des classes d'affixes : l'idée étant de regrouper des stemmes (respectivement des affixes) qui sont attestés avec les mêmes affixes (respectivement stemmes). Le critère optimisé pour construire ces regroupements est un critère de type *Minimum Description Length.*

Segmentation morphématique Une troisième sous-tâche ressortissant de l'analyse morphologique consiste à segmenter des formes en unités plus petites, correspondant idéalement aux morphèmes[1] de la tradition classique en morphologie (*Item and arrangements*, voir [Fradin, 2003] pour une introduction critique à ce

[1] Ou plutôt aux allomorphes : la segmentation porte le plus souvent sur les formes de surface.

modèle). Construire une telle segmentation présente un intérêt par exemple pour des applications de conversion orthographique-phonétique, les frontières morphématiques internes pouvant avoir une influence sur la prononciation d'une forme (ainsi pour les prononciations d'*asocial* ou de *parasol* en français), ou encore sur la position de l'accent lexical. Une autre application possible, mentionnée par [Creutz and Lagus, 2002], est la construction de modèles stochastiques des séquences de sous-mots, utilisables pour des applications de reconnaissance de la parole. [van den Bosch et al., 1996] développe une stratégie d'apprentissage supervisé pour cette tâche, étendue dans [van den Bosch and Daelemans, 1999] par la prédiction d'une étiquette catégorielle pour les frontières de morphèmes. Considérée sous cet angle, l'analyse morphologique peut être modélisée et traitée comme un problème de catégorisation binaire des éléments d'une séquence.

Liste de morphèmes La majorité des travaux porte toutefois sur l'apprentissage *non-supervisé* de stratégies de segmentation, un produit dérivé (ou final) étant l'identification de listes de morphèmes, voire d'un modèle simpliste de la morphotactique. L'approche dominante sur cette question, représentée e.g. par [de Marcken, 1996, Deligne et al., 1996, Déjean, 1998] s'inspire plus ou moins explicitement des travaux originels de Harris [1970], qui suggère d'utiliser une mesure informationnelle pour déterminer les frontières de morphèmes. L'intuition principale d'Harris est que l'expression de la combinatoire morphématique se lit dans les distributions conditionnelles des lettres à l'intérieur des formes. En suivant les évolutions de l'entropie conditionnelle de la distribution des lettres au fur et à mesure que l'on avance dans un mot, il est possible de détecter les frontières de morphèmes, qui correspondent à des pics d'entropie, c'est-à-dire à des zones où les probabilités de recombinaison sont les plus fortes. [Monson et al., 2004] revient sur cette stratégie de découverte de morphèmes, en la couplant à l'induction de classes de suffixes opérant sur des stemmes communs. [Snover and Brent, 2001] vise également à apprendre de manière non supervisée un ensemble de suffixes ; ce travail se distingue des précédents par l'utilisation d'un modèle probabiliste complet de génération du lexique, qui permet de comparer différentes segmentations au bénéfice de la segmentation la plus probable.

A.0.8 Analyse syntaxique : petit à petit...

La tradition classique en traitement automatique des langues envisage essentiellement la tâche d'analyse syntaxique de manière globale, comme la construction d'une (ou plusieurs) structure(s) représentant un énoncé naturel. Cette étape de traitement constitue un préalable au calcul d'une représentation du sens de l'énoncé. La structure produite prend, selon les modèles formels privilégiés par telle ou telle théorie linguistique, la forme d'un arbre étiqueté, d'une structure de traits (éventuellement typée), d'une preuve logique, ou encore combine plusieurs de ces éléments [Abeillé, 1993]. Il est usuel d'exiger qu'une telle représentation, pour valoir comme analyse syntaxique, encode l'organisation de l'énoncé en constituants et sous-constituants, ainsi que les relations de dépendance et de domination entre constituants, voire parfois également leur fonction syntaxique.

La mise en œuvre d'une telle approche conduit à expliciter un ensemble de *règles* (par exemple des règles de récriture) ou de *contraintes*, associées à un dispositif algorithmique permettant de calculer effectivement cette association. Un produit dérivé de cette méthodologie est l'identification des énoncés « agrammaticaux », qui sont ceux pour lesquels le calcul d'une représentation échoue.

Le renouveau des techniques à base de corpus a conduit à modifier sensiblement cette conception : en particulier, il est apparu que l'exploitation de corpus pouvait utilement servir de support à une conception plus graduelle (exemplairement probabilisée) de la notion de grammaticalité. Cette réflexion a nourri le développement de modèles visant à assigner un score de vraisemblance à un énoncé, ou aux analyses d'un énoncé. De tels modèles se sont avérés fournir des outils puissants pour de nombreuses tâches de *désambiguïsation*, que cette désambiguïsation ait pour but de départager plusieurs énoncés en compétition (cas de la reconnaissance vocale, de la reconnaissance de documents numérisés, de la correction orthographique...) ou de

départager plusieurs analyses également valides d'un énoncé syntaxiquement ambigu.

Les plus simples des modèles probabilistes de la syntaxe, connus sous le nom de *modèles de langage*, sont très succintement introduits et discutés ci-dessous. Il existe également une abondante littérature portant sur l'apprentissage de modèles plus complexes, tels que, par exemple, les grammaires hors-contexte probabilistes (voir e.g. [Charniak, 1993, Bod, 1995, Abney, 1997, Goodman, 1998, Collins, 1999, Roark, 2001]). Je n'en dirai pas plus ici.

Parallèlement, sous la pression d'une demande croissante pour des analyseurs syntaxiques moins précis dans leurs analyses, mais capables de couvrir un plus large ensemble d'énoncés, une autre gamme d'analyseurs a progressivement émergé, qui envisage le processus d'analyse comme un processus itératif, au cours duquel la structure se construit par enrichissement progressif de l'énoncé : étiquetage morpho-syntaxique, identification de « pseudo-syntagmes » (*chunks*), désambiguïsation des rattachements et des liens de dépendance, détection et suivi des chaînes de coréférence, etc. Cette approche est particulièrement adaptée lorsqu'une analyse complète n'est en fait pas nécessaire. Ainsi, par exemple, en synthèse vocale à partir du texte, il est courant de s'en tenir à la construction d'une segmentation en pseudo-constituants, qui suffit en général pour calculer la prosodie des énoncés. Je m'en en tiendrai ici à la présentation de la plus emblématique de ces tâches, à savoir l'étiquetage morpho-syntaxique (section A.0.8), en signalant toutefois que les autres tâches ont également donné lieu à l'utilisation de techniques d'apprentissage, en particulier les tâches de chunking (qui est une instance de tâche d'étiquetage, voir 1.2.3) et de désambiguïsation des liens de dépendance et des chaînes de coréférence, qui sont susceptibles, moyennant une reformulation idoine, d'être traitées par des techniques de catégorisation supervisée.

Modèles de langage

La modélisation statistique du langage [Jelinek, 1997] vise à assigner un score de vraisemblance à toute séquence finie de mots pris dans un ensemble fini V. Comme discuté ci-dessus, de tels modèles jouent un rôle crucial dans de nombreuses applications, en particulier dans le cadre de la reconnaissance vocale (voir également la section 4.2).

Si l'on note $w = w_1^T$ une séquence de mots, l'approche la plus simple, mise en œuvre dans les modèles de langage de type k-gram, modélise $P(w_1^T)$ selon :

$$P(w) = P(w_1)P(w_2 \mid w_1)\ldots\prod_{t=k}^{t=T} P(w_t \mid w_{t-1}\ldots w_1) \quad \text{(A.1)}$$

$$\approx P(w_1)P(w_2|w_1)\ldots\prod_{t=k}^{t=T} P(w_t|w_{t-1}\ldots w_{t-k+1}) \quad \text{(A.2)}$$

Le passage de (A.1) à (A.2) consiste à considérer que la probabilité d'occurrence du t-ème terme de la séquence, conditionnellement à son historique $h_t = w_1 \ldots w_{t-1}$, ne dépend que des $k-1$ mots précédents. Sous cette hypothèse, le modèle comprend un nombre fini de paramètres : pour chaque valeur possible de h (il en existe $|V|^{k-1}$ pour un modèle k-gramme), ces paramètres correspondent aux V valeurs $P(w|h)$ des lois discrètes $P(.|h)$.

L'estimation au maximum de vraisemblance de ces distributions conduit à calculer sur un corpus les fréquences relatives des séquences hw, rapportées à celle de la séquence h. Même pour des petites valeurs de k, telles que celles utilisées en pratique ($k = 2$ ou $k = 3$), le nombre de paramètres, conjoint à la très inégale distribution des séquences hw dans les corpus, rend l'estimation de tels modèles extrêmement problématique. En effet, même si l'on utilise de très larges corpus textuels, l'immense majorité des séquences hw ne sera jamais observée, sans que l'on puisse facilement distinguer les paires qui sont quasi-impossibles, de celles dont l'absence dans le corpus n'est que fortuite. Dans ce contexte, l'utilisation de techniques de

lissage, permettant d'assigner des probabilités non-nulles à des observations de fréquence nulle est indispensable. Ces techniques s'appuient pour l'essentiel sur la combinaison de modèles allant du plus au moins complexe : la probabilité $P(w|h)$ se dérivant de la probabilité $P(w|\overline{h})$, où $\overline{h}$ est l'historique *tronqué*.

Les modèles k-gramme présentent un défaut majeur : ils ne permettent de prendre en compte que des dépendances locales : chaque mot est prédit par ses voisins proches. Pour intégrer des dépendances plus longues, les propositions les plus convainquantes restent celles formulées dans [Rosenfeld, 1994, 2000] dans le cadre de modèles du maximum d'entropie. Le premier de ces travaux propose d'exprimer la probabilité conditionnelle $P(w|h)$ sous la forme :

$$P(w|h) = \frac{1}{Z(h)} \exp(\sum_k \lambda_k f_k(w,h)) \tag{A.3}$$

où chaque f_k est une fonction booléenne du mot courant et de l'historique, incluant potentiellement des dépendances longues, telles que par exemple l'apparition de w ou d'un terme sémantiquement apparenté dans une fenêtre de passé large, pouvant aller au-delà des frontières de phrase. Deux difficultés se posent toutefois, pour lequelles Rosenfeld étudie diverses solutions : la détermination des fonctions les plus informatives et l'estimation des paramètres.

La seconde proposition étend cette approche en proposant des modèles exponentiels de phrase selon :

$$P(w) = \frac{1}{Z} \exp(\sum_k \lambda_k f_k(w)) \tag{A.4}$$

où les f_k testent des propriétés arbitraires de l'énoncé : la fréquence de bigrammes ou de tri-grammes, mais également des configurations syntaxiques, la cohérence sémantique, etc.

L'étiquetage morpho-syntaxique

La tâche d'étiquetage morpho-syntaxique consiste à assigner à chaque forme d'un texte une étiquette correspondant à la catégorie syntaxique principale de la forme en contexte. L'information catégorielle est éventuellement complétée par des informations morpho-syntaxiques supplémentaires (genre, nombre, personne, temps, mode...). Un exemple de suite de couples entrée-sortie, extrait des données de l'action GRACE [Adda et al., 1998], est donné dans la table A.2.

Forme	Étiquette	Glose
le	Da-ms-d	Déterminant masculin singulier défini
père	Ncms	Nom commun masculin singulier
Mathieu	Npms	Nom propre masculin singulier
avait	Vaii3s-	Verbe auxiliaire indicatif imparfait 3eme personne singulier
eu	Vmps-sm	Verbe principal participe passé masculin singulier
un	Da-ms-i	Déterminant, masculin, singulier, indéfini
mouvement	Ncms	Nom commun masculin singulier
violent	Afpms	Adjectif qualificatif positif masculin singulier

TAB. A.2 – L'entrée et la sortie d'un système d'étiquetage morpho-syntaxique

Suivant les applications envisagées, le nombre d´étiquettes peut varier de quelques dizaines à quelques centaines. La mise en œuvre d'outils de désambiguïsation répond au constat de la prolifération des ambiguïtés lexicales : en français, il est couramment admis qu'environ la moitié des formes apparaissant dans les textes peut être rattachée à plusieurs lexèmes. Pour donner un seul exemple, *bus* est à la fois une forme (ambiguë pour le nombre) du lexème BUS et une forme (ambiguë pour la personne et pour le temps) du lexème BOIRE. On note toutefois que, pour une forme donnée, tous les lexèmes ne sont pas équiprobables : *la* est beaucoup plus souvent une forme du déterminant ou du pronom personnel qu'un nom désignant une note de musique.

Les descriptions grammaticales fondées sur des grammaires formelles (CFG, LFG, HPSG, LTAG...) qui intègrent le concept de catégorie syntaxique doivent faire face, lors de l'analyse, à un problème majeur de gestion des ambiguïtés. Ces ambiguïtés sont, de surcroît, susceptibles de se propager localement et de ralentir considérablement l'analyse, quand elles ne la vouent pas à l'échec.

Dès le milieu des années 80[2], [Church, 1988, DeRose, 1988] proposent de considèrer cette étape de désambiguïsation morpho-syntaxique comme une tâche séparée, préalable à l'analyse syntaxique et montrent que cette nouvelle tâche peut être effectuée efficacement par des techniques statistiques (dans ces travaux, il s'agit de HMMs). Ces deux articles soulignent également que cette tâche est susceptible de trouver d'autres applications que l'analyse syntaxique, par exemple en synthèse de la parole à partir du texte, ou encore pour préparer la lemmatisation. Enfin, cette tâche peut donner lieu à des évaluations objectives (numériques) et à des comparaisons entre systèmes.

Ces travaux ont donné naissance à une importante littérature et à de nombreuses implantations de systèmes utilisant des techniques d'apprentissage pour la désambiguïsation morpho-syntaxique (en particulier [Brill, 1993]) ; comme elle a encouragé la mise au point de corpus étiquetés et la mise en place de méthodologies d'évaluation objectives. Ces travaux ultérieurs ont permis de mettre en évidence l'importance de prédire correctement les étiquettes possibles pour un mot inconnu (problème évoqué ci-dessus) et de relativiser le succès des approches statistiques par comparaisons avec les méthodes à base de règles d'experts [Chanod, 1995, Samuelsson and Voutilainen, 1997] (voir également, pour le français, les travaux conduits dans le cadre de l'action GRACE [Adda et al., 1998]).

De manière plus générale, cette tâche a servi de modèle à de nombreuses autres tâches, qui ont pu être reformulées dans les mêmes termes d'une désambiguïsation en contexte, reconsidérées avec les mêmes outils statistiques, et également soumises à une évaluation « objective ». Dans la brève histoire du domaine, on peut considérer qu'elle a servi de déclencheur pour réhabiliter les approches statistiques en syntaxe et plus généralement dans tout le champ du traitement automatique des langues.

A.0.9 Sémantique

L'analyse sémantique d'un énoncé vise à construire une représentation formelle de son sens littéral (par exemple sous la forme d'une formule logique), en s'appuyant sur les informations collectées durant les étapes de structuration précédentes, ainsi que sur des ressources sémantiques, notamment lexicales.

Cet objectif reste encore aujourd'hui hors d'atteinte pour des systèmes visant une large couverture, faute notamment de disposer des ressources sématiques nécessaires à l'achèvement de ce but. Les travaux actuellement conduits en sémantique se divisent alors en deux catégories principales : d'une part ceux qui s'appliquent à apprendre automatiquement des ressources sémantiques lexicales à partir de corpus, utilisant en particulier des méthodes d'apprentissage non-supervisé (j'en ai évoqué quelques-uns à la section 1.2.1, voir également mes propres travaux, décrits à la section 3.1) ; d'autre par ceux qui cherchent à accomplir des sous-tâches plus simples, qui apparaissent comme des prérequis à la construction d'une représentation sémantique complète : identification d'unités linguistiques ayant une sémantique stable (montants, noms propres, etc.) ; désambiguïsation du sens des formes polysémiques ; détection de relations d'implications, etc. L'intérêt de ces différents traitements étant qu'ils sont d'ores et déjà susceptibles de servir des applications de gestion documentaire (indexation, recherche, etc.) ou encore de traduction.

Les paragraphes qui suivent introduisent brièvement les deux premières de ces tâches ; on se reportera notamment à [Dagan et al., 2005] pour les travaux récents concernant le calcul de relations d'implications.

[2] [Church, 1988] fait référence à des travaux antérieurs, conduits en particulier à IBM [Bahl and Mercer, 1976] et à l'université de Lancaster [Leech et al., 1983].

Repérage des entités nommées

La tâche de repérage des entités nommées dérive des travaux en *extraction d'information* [Poibeau, 2003] : pour un ensemble homogène de textes, rapportant des informations de nature similaire, il s'agissait à l'origine, de repérer les éléments textuels susceptibles de remplir les différents champs d'un formulaire résumant le texte. Par exemple, pour le domaine des incidents terroristes [MUC, 1991], le type d'information recherché au sein des documents comprend la date, le lieu, l'identité du ou des auteur(s), le nombre de victimes, etc.

Il est progressivement apparu qu'un composant essentiel des systèmes d'extraction d'information était leur capacité à repérer les éléments textuels sémantiquement stables[3], tels que les noms de personnes ou de lieux, les dates, montants, etc., qui constituent des éléments d'information essentiels pour le bon remplissage des formulaires. Cette (sous)-tâche a, en conséquence, fait l'objet de multiples campagnes d'évaluation, sur des textes dans le cadre des campagnes MUC [Chinchor, 1996, Chinchor and Robinson, 1997], sur des transcriptions d'émissions radio-télévisées, ainsi, que, plus récemment, sur des corpus multilingues dans le cycle des conférences CoNLL [Tjong Kim Sang, 2002, Tjong Kim Sang and de Meulder, 2003] et sur des corpus bio-médicaux [Collier et al., 2004].

D'un point de vue pratique, le repérage des entités nommées consiste à introduire des balises dans un texte autour des fragments qui constituent des entités nommées pour le domaine considéré, en précisant, pour chaque entité, le type de référence qu'elle désigne[4]. Selon les domaines, les types valides peuvent correspondre effectivement à des noms propres (de personnes, d'institution, de lieux, de composés chimiques, de gènes...), à des dates, à des montants, etc. Deux représentations alternatives pour cette tâche, tirées d'une évaluation conduite sur des données audio, sont représentées à la figure A.1. Dans la première représentation, les entitées sont simplement encapsulées dans des balises ; dans la seconde, chaque mot est étiqueté comme étant soit une frontière de début (B), soit une continuation (I), soit à l'extérieur d'une entité (O). Cette seconde représentation permet l'utilisation directe de techniques de catégorisation supervisée.

```
<B_ENAMEX TYPE="PERSON">DAVID WHITCOMB<E_ENAMEX> IS A
<B_ENAMEX TYPE="ORGANIZATION">RUTGERS UNIVERSITY<E_ENAMEX>
FINANCE PROFESSOR AND FREQUENT
<B_ENAMEX TYPE="ORGANIZATION">NASDAQ<E_ENAMEX> CRITIC.
```

DAVID	B-PERSON
WHITCOMB	I-PERSON
IS	O
A	O
RUTGERS	B-ORGANIZATION
UNIVERSITY	I-ORGANIZATION
FINANCE	O
...	...

FIG. A.1 – Deux représentations pour la tâche de repérage des entités nommées

Compte-tenu de son importance pour des applications d'extraction d'information, mais également pour les

[3] La règle générale permettant de décider si une expression est ou n'est pas une entité nommée se fonde sur le caractère d'unicité du concept représenté par l'expression. Les inévitables ambiguïtés produites par une définition aussi générale sont traitées par un ensemble de recommandations à l'usage des annotateurs de corpus [Chinchor and Robinson, 1997].

[4] De ce point de vue, la tâche de repérage des entités nommées vaut plus largement comme représentant d'une vaste gamme de tâches consistant à identifier (et à typer) des segments dans un flux textuel et dont ressortissent également de nombreuses sous-tâches de l'analyse syntaxique : segmentation en chunk, identification des groupes nominaux ou verbaux, etc.

systèmes d'indexation, de question-réponse..., l'identification des entités nommées a fait l'objet de travaux nombreux. L'approche « linguistique » pour ce problème consiste à définir, pour chaque type d'entité nommée, (i) des dictionnaires d'entités nommées, (ii) des expressions rationnelles décrivant la forme typique que peuvent prendre les entités et (iii) les contextes syntaxiques usuels dans lesquelles elles apparaissent [Appelt et al., 1993]. Avec la disponibilité croissante de corpus annotés, cette tâche a également donné lieu à de très nombreux travaux utilisant des techniques d'apprentissage automatique (voir, par exemple, [Cardie, 1997, Bikel et al., 1999, Tjong Kim Sang, 2002, Tjong Kim Sang and de Meulder, 2003]), qui pour l'essentiel, utilisent les modèles de transcodage introduits à la section 1.1.2.

Désambiguïsation sémantique

La tâche de désambiguïsation sémantique est historiquement liée aux applications de traduction automatique et vise à s'attaquer au problème de la polysémie lexicale. De nombreux lexèmes ont, en effet, plusieurs sens différents, qui vont donner lieu à des traductions différentes. C'est le cas par exemple en anglais du nom *bank*, qui se traduira suivant les contextes *banque* ou *rive*. Désambiguïser le sens des mots a de multiples autres applications potentielles : ainsi, dans toutes les tâches qui relèvent de la fouille de texte (voir ci-dessous), il semble qu'un tel traitement pourrait améliorer la qualité de l'indexation des documents. On se reportera à [Ide and Véronis, 1998], pour une discussion de la tâche désambiguïsation sémantique, depuis les débuts de la traduction automatique jusqu'aux travaux récents. Le cycle d'évaluation Senseval[5] permet d'apprécier l'ampleur des efforts accomplis sur cette tâche et les performances actuelles des systèmes représentatifs de l'état de l'art.

L'établissement de la désambiguïsation sémantique comme une tâche autonome nécessite que soit définie la sortie du système, et donc de décider, pour chaque lexème, quels sont les différents sens qu'il s'agit de discriminer. L'établissement d'une telle liste n'est pas sans difficulté et dépend en réalité des applications. Ainsi, par exemple, la distinction entre les deux sens de *river* est sans doute nécessaire pour une application de traduction pour des textes géographiques[6], mais semble moins nécessaire pour des tâches d'indexation.

Dans la pratique, on a couramment recours à des dictionnaires (mono-lingues ou bilingues) ou, lorsqu'elles existent, à des ressources sémantiques de type Wordnet. Les deux types de ressources ont leurs limitations : en ce qui concerne les dictionnaires bilingues, notons ainsi qu'une différence de traduction n'implique pas toujours une différence de sens : un terme non-ambigu peut avoir des traductions différentes dans une autre langue. Ainsi, *éteindre* en français, dont la traduction en anglais dépend de la nature de l'objet éteint (*to turn off* pour un appareil électrique, *to blow* pour une chandelle, *to stub out* pour une cigarette, etc). L'inverse est également possible, puisqu'il existe des termes dont l'ambiguïté de sens est conservée dans les deux langues : c'est le cas du couple (*souris*/*mouse*), pour lequel l'ambiguïté entre l'animal et le périphérique informatique existe en anglais comme en français. Les ressources comme Wordnet ont également leurs défauts : elle vont ainsi enregistrer des différences extrêmement fines pour certains termes ou domaines, mais pas dans d'autres (ainsi, *river* n'a qu'un sens dans Wordnet). On peut globalement résumer ces difficultés en réalisant qu'il existe ainsi toute une gamme de situations qui vont de la polysémie entièrement accidentelle (en synchronie, où l'on peut considérer qu'on a affaire à des lexèmes différents) à des variations de sens subtiles d'un même lexème [Victorri and Fuchs, 1996].

Si l'on fait toutefois l'hypothèse qu'une telle liste de sens puisse être établie, la désambiguïsation sémantique se présente comme un pur problème de catégorisation, puisqu'il s'agit d'affecter à chaque occurrence en contexte d'un terme sémantiquement ambigu une étiquette qui correspond au sens retenu pour cette occurrence. La désambiguïsation peut s'appuyer sur la catégorie syntaxique du terme, lorsqu'elle est connue, ainsi que sur son contexte local (*micro-contexte* : au plus la phrase) ou global (allant jusqu'au document

[5] `http://www.senseval.org`.

[6] *fleuve* si le cours d'eau se jette dans la mer et par *rivière* sinon.

complet) d'occurrence ([Preiss, 2004] est un exemple de système de désambiguïsation sémantique s'appuyant des informations contextuelles variées).

La difficulté principale pour réaliser un tel apprentissage provient de la rareté de la ressource d'apprentissage (des corpus désambiguïsés en sens), difficulté qu'amplifie la très inégale distribution des sens : certains pouvant être marginaux dans la langue générale mais dominants pour des langues de spécialité. Un effort tout particulier en désambiguïsation sémantique a donc porté sur les techniques alliant apprentissage supervisé et non-supervisé (construction de classes de mots, techniques de *co-training*, etc.).

Bibliographie

Proceedings of the third Message Understanding Conference (MUC-3), San Francisco, CA, 1991. Morgan Kaufman Publishers.

Anne Abeillé. *Les nouvelles syntaxes*. Armand Colin, Paris, 1993.

Steve Abney. Stochastic attribute-value grammars. *Computational Linguistics*, 23(4) :597–618, 1997.

Gilles Adda, Josette Lecomte, Jospeh Mariani, Patrick Paroubek, and Martin Rajman. The GRACE French part-of-speech tagging evaluation task. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain, 1998.

Evan L. Antworth. *PC-KIMMO : a two-level processor for morphological analysis*. Number 16 in Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas,TX, 1990. ISBN 0-88312-639-7.

Douglas Appelt, Jerry Hobbs, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. SRI : Description of the JV-FASTUS system used for muc-5. In Beth Sundheim, editor, *the Proceedings of the Fifth Message Understanding Conference (MUC-5)*, Tokyo, Japan, 1993.

Lalit R. Bahl and Robert L. Mercer. Part of speech assignment by a statistical decision algorithm. In *IEEE International Symposium on Information Theory*, pages 88–89, Ronneby, 1976.

Marco Baroni, Johannes Matiasek, and Harald Trost. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceeding of the ACL Workshop on Morphological and Phonological Learning*, Philadelphia, PA, 2002.

Frédéric Béchet, Thierry Spriet, and Marc El-Bèze. Automatic assignment of part-of-speech to out-of-vocabulary words for text-to-speech processing. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece, 1997.

Daniel M. Bikel, Richard Schwartz, and Ralph Weischedel. An algorithm that learns what's in a name. *Machine Learning : Special Issue on Natural Language learning*, 34(1–3) :211–231, 1999.

Maximilian Bisani and Hermann Ney. Multigram-based grapheme-to-phoneme conversion for lvcsr. In *INTERSP*, pages 933–936, Geneva, Switzerland, 2003.

Rens Bod. *Enriching Linguistics with Statistics : Performance Models of Natural Language*. Institute for Logic, Language and Computation, Amsterdam, 1995.

Eric Brill. Automatic grammar induction and parsing free text : a transformation-based approach. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 259–265, Columbus, Ohio, 1993.

Eric Brill. Some advances in rule based part of speech tagging. In *Proceedings of AAAI*, 1994.

Claire Cardie. Empirical methods in information extraction. *AI magazine*, 18(4) :65–80, 1997.

Jean-Pierre Chanod. Tagging French : comparing a statistical and a constraint-based method. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 149–156, Dublin, Ireland, 1995.

Eugene Charniak. *Statistical Language Learning*. The MIT Press, Cambridge, MA, 1993.

Stanley F. Chen and Benoît Maison. Using place name data to train language identification models. In *INTERSP*, pages 1349–1353, Geneva, Switzerland, 2003.

Nancy Chinchor. Muc-6 named entity task definition. In *Proceedings of the sixth Message Understanding Conference (MUC-6)*, San Francisco, CA, 1996.

Nancy Chinchor and Patricia Robinson. Muc-7 named entity task definition. In *Proceedings of the seventh Message Understanding Conference (MUC 7)*, 1997.

Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied NLP*, pages 136–143, Austin, Texas, 1988.

Nigel Collier, Patrick Bush, and Adeline Nazarenko, editors. *Proceedings of the joint international workshop on natural language processing in biomedicine and its application*, Genève, Suisse, 2004. URL http://www.genisis.ch/~natlang/JNLPBA04/program.html.

Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proceeding of the ACL Workshop on Morphological and Phonological Learning*, page xxx, Philadelphia, PA, 2002.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

Carl de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Dpt of Computer Science, MIT, 1996.

Hervé Déjean. Morphemes as necessary concept for structures discovery from untagged corpora. In *Proceedings of the Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299, Adelaide, Australia, 1998.

Sabine Deligne, François Yvon, and Frédéric Bimbot. Introducing statistical dependencies and structural constraints in variable-length sequence models. In Laurent Miclet and Colin de la Higuera, editors, *Grammatical Inference : Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence 1147, pages 156–167. Springer, September 1996.

Steven J. DeRose. Grammatical category desambiguation by statistical optimization. *Computational Linguistics*, 14(1) :31–39, 1988.

Michel Divay and Anthony J. Vitale. Algorithm for grapheme-to-phoneme translation for French and English : Applications. *Computational Linguistics*, 23(4) :495–524, 1997.

Steve Finch and Nick Chater. Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, pages 820–825, 1992.

Bernard Fradin. *Nouvelles approches en morphologie*. Presses Universitaires de France, Paris, France, 2003.

Eric Gaussier. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Methods in Natural Language Processing*, College Park, MD, 1999.

John Goldsmith. Unsupervised learning of the morphology of natural languages. *Computational Linguistics*, 27(2) :153–198, 2001.

Joshua Goodman. *Parsing Inside-Outside*. PhD thesis, Division of Engineering and Applied Sciences, Harvard University, 1998.

Zellig Harris. *Papers in Structural and Transformational linguistics*. D. Reidel, Dordrecht, 1970.

Nancy Ide and Jean Véronis. Introduction to the special issue on word sense desambiguation. *Computational Linguistics*, 24(1) :1–40, 1998.

Christian Jacquemin. Guessing morphology from terms and corpora. In *Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 156–167, Philadelphia, PA, 1997.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, MA, 1997.

Kimmo Koskenniemi. *Two-level morphology : a general computational model for word-form recognition and production*. Number 11 in Publication. University of Helsinki Department of General Linguistics, Helsinki, 1983.

Bob Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, 1993.

Geoffrey Leech, Roger Garside, and Eric Atwell. The automatic grammatical tagging of the LOB corpus. *ICAME News*, (7) :13–33, 1983.

Andrei Mikheev. Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3) : 405–423, 1997.

Christian Monson, Alon Lavie, Jaime Carbonell, and Lori Levin. Unsupervised induction of natural language morphology inflection classes. In *In the Proceedings of the Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, Barcelona, 2004.

Thierry Poibeau. *Extraction automatique d'information, du texte brut au web sémantique*. Hermès, Paris, France, 2003.

M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3) :130–137, 1980.

Judita Preiss. Probabilistic word sense disambiguation. *Computer Speech and Language*, 18 :319–337, 2004.

Ardwait Rathnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, 1998.

Brian Roark. *Robust Probabilistic Predictive Syntactic Processing : Motivations, Models and Applications*. PhD thesis, Dpt. of cognitive and linguistic sciences, Brown University, 2001.

Ronald Rosenfeld. Two decades of statistical language modeling : Where do we go from here ? *Proceedings of the IEEE*, 88(8), 2000.

Ronald Rosenfeld. *Adaptative Statistical Language Modeling : A Maximimum Entropy Approach*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1994.

Christer Samuelsson and Atro Voutilainen. Comparing a linguistic and a stochastic tagger. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 246–253, Madrid, Spain, 1997.

Patrick Schone and Dan Jurafsky. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, Lisbon, Portugal, 2000.

Patrick Schone and Dan Jurafsky. Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, Pittsburgh, PA, 2001.

Matthew G. Snover and Michael R. Brent. A bayesian model for morphemes and paradigm identification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 482–490, Toulouse, France, 2001.

Richard Sproat. *Morphology and Computation*. The MIT Press, 1992.

Richard Sproat, Alan W. Black, Stanley F. Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. Normalization of non-standard words. *Computer Speech and Language*, 15 :287–333, 2001.

Richard W. Sproat, editor. *Multi-lingual Text-to-Speech Synthesis : the Bell labs apprach*. Kluwer Academic Publishers, Boston, MA, 1997.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task : Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan, 2002. URL http://cnts.uia.ac.be/conll2002/proceedings.html.

Erik F. Tjong Kim Sang and Fien de Meulder. Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 155–158. Edmonton, Canada, 2003. URL http://cnts.uia.ac.be/conll2003/proceedings.html.

Antal van den Bosch and Walter Daelemans. Memory-based morphological processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 285–292, College Park, MD, 1999.

Antal van den Bosch, Walter Daelemans, and Tom Weijters. Morphological analysis as classification : an inductive-learning approach. In Kemal Oflazer and Harald Somers, editors, *Proceedings of the Second International Conference on New Methods in Language Processing (NEMLAP II)*, pages 79–89, Ankara, Turkey, 1996.

Bernard Victorri and Catherine Fuchs. *La polysémie, construction dynamique du sens*. Hermès, Paris, France, 1996.

François Yvon. *Prononcer par analogie : motivations, formalisations et évaluations*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 1996.

Rémy Zajac. Morpholog : Constrained and supervised learning of morphology. In *Proceedings of the conference on Computational Natural Language Learning (CoNLL-2001)*, Toulouse, France, 2001.

Annexe B

Sélection d'articles

Articles relatifs au chapitre 2

Nicolas Stroppa and François Yvon (2005). An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pages 120–127, Ann Arbor, MI. (8 pages)

Nicolas Stroppa and François Yvon (2005). Analogical learning and formal proportions : Definitions and methodological issues. Technical Report ENST-2005D004, École Nationale Supérieure des Télécommunications, Paris, France, 2005. (14 pages)

Articles relatifs au chapitre 3

Loïs Rigouste, Olivier Cappé, and François Yvon (2006). Inference and Evaluation of the Multinomial Mixture Model for Text Clustering. Technical Report ENST-2006D004, École Nationale Supérieure des Télécommunications, Paris, France, 2006. (26 pages)

Romain Vinot and François Yvon (2003). Improving Rocchio with weakly supervised clustering. In the *Proceedings of the European Conference on Machine Learning (ECML)*, pages 456–467, Cavtat-Drubrovnick, Croatie, 2003. (8 pages)

Articles relatifs au chapitre 4

Shiuan-Sung Lin and François Yvon (2005). Discriminative training of finite-state decoding graphs. In the *Proceedings of the International Conference on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal. (4 pages)

François Yvon, Geoffrey Zweig, and George Saon (2004). Arc minimization in finite-state decoding graphs with cross-word acoustic context.*Computer Speech and Language*, 18(4) :397–415. (24 pages)