Modèles multi-thématiques markoviens pour la segmentation de textes

Loïs Rigouste*, Olivier Cappé*, François Yvon*
et Fabrice Clérot**

* GET – Télécom Paris & CNRS – LTCI
46 rue Barrault, 75634 Paris Cédex 13
rigouste,cappe,yvon à enst.fr

** France Télécom Division R & D TECH/SUSI/TSI
2 Avenue Pierre Marzin, 22307 Lannion Cédex
fabrice.clerot à francetelecom.com

Résumé. Dans cet article, nous montrons comment des outils génériques de la fouille statistique de textes peuvent être utilisés pour résoudre une tâche d'apprentissage supervisée: le DÉfi Fouille de Textes 2005. Dans un premier temps, nous étudions comment capturer une partie des spécificités de la tâche à l'aide de modèles de Markov cachés. Nous détaillons ensuite une modélisation des textes par un mélange de distributions multinomiales sur les comptes de mots, dans laquelle chaque composante correspond à un thème particulier. Les paramètres des distributions thématiques sont estimés grâce à l'algorithme EM. Ce modèle est utilisé pour diviser en sous-thèmes les discours des deux présidents. Nous discutons finalement des performances obtenues en combinant ces deux outils.

1 Introduction

La tâche DEFT, introduite plus en détail dans ce même numéro, consiste à analyser un pseudo-document construit en insérant, dans un discours de Jacques Chirac, un fragment de discours de François Mitterrand. Il s'agit, pour les participants, de séparer le document original de l'insert éventuel. Ils peuvent, à cette fin, s'appuyer sur un corpus de pseudo-documents annotés par les organisateurs.

Cette tâche se prête *a priori* à plusieurs approches :

– l'identification *non-supervisée* de segments thématiquement homogènes dans les pseudo-documents, problème pour lequel de multiples méthodes sont disponibles (voir, par exemple, (Hearst, 1997; Choi, 2000)), qui toutes essaient de tirer parti de l'organisation séquentielle du texte. Cette démarche est confortée par la méthodologie de constitution de la base de données, selon laquelle les insertions de discours de François Mitterrand traitent de thématiques différentes de celles des discours de Jacques Chirac. Cette stratégie a pour inconvénient de ne pas réellement exploiter les données de supervision disponibles.

– la classification *supervisée* des phrases dans deux catégories, une pour chaque président. Cette tâche est bien documentée dans la littérature et il existe de nombreux outils permettant de la résoudre (voir, par exemple, (Sebastiani, 2002) pour une revue). Cette approche, qui permet de tirer effectivement parti des données de supervision, se heurte à une double difficulté. D'une part, les fragments à catégoriser sont courts (des phrases isolées), ce qui fragilise les systèmes de catégorisation ; d'autre part, on peut s'attendre à ce que les deux classes soient « proches », dans la mesure où, d'un point de vue purement thématique, les échantillons de phrases des deux présidents abordent globalement des sujets très voisins. Une variante, permettant de répondre partiellement à la seconde objection consisterait à fonder la discrimination sur des attributs purement stylistiques (longueur des phrases, fréquence d'emplois de certains marqueurs lexicaux ou syntaxiques), en faisant abstraction des mots sémantiquement pleins. Cette approche est caractéristique des travaux portant sur l'identification d'auteurs en lexicométrie (Benzécri et al., 1981).

La plupart des participants à DEFT'05 ont proposé des méthodes empruntant à ces deux démarches. La cohérence de la segmentation est assurée soit par des modèles de Markov cachés (Hidden Markov Models, HMM) (Jelinek, 1997), soit par les outils de segmentation cités plus haut, alors que la tâche de classification supervisée fait l'objet de traitements plus variés : classifieurs bayésiens avec extraction de divers attributs (El-Bèze et al., 2005; Labadié et al., 2005), analyse de dépendances syntaxiques (Maisonnasse et Tambellini, 2005) ou machines à vecteurs supports (Kerloch et Gallinari, 2005) sont des exemples parmi d'autres.

L'approche que nous avons mise en œuvre suit également cette idée de combinaison de segmentation et de classification, en exploitant des outils *génériques* de la fouille de texte, réutilisables dans de nombreux autres contextes. Elle utilise, d'une part, des techniques de catégorisation et de classification probabilistes, qui se fondent sur une représentation en « sac-de-mots » des phrases. Ainsi, la phase d'apprentissage consiste à inférer les paramètres de modèles multi-thématiques des discours de chaque président. Notre approche vise, d'autre part, à tirer profit des différentes contraintes de la tâche, et en particulier celles qui se déduisent de la méthode de constitution des pseudo-documents, qui sont exprimées par des automates finis. L'ensemble est combiné sous la forme de HMMs. La figure 1 illustre l'approche dans sa globalité.

Cet article est organisé comme suit. Nous décrivons à la section 2 le principe général de l'algorithme de segmentation, qui repose essentiellement sur la technologie des HMMs. Nous montrons comment, à partir d'un modèle simple, il est possible d'intégrer progressivement les différentes contraintes de la tâche en utilisant des structures de plus en plus élaborées. Cette section se termine par l'adaptation des HMMs au cas où les discours de chaque président peuvent être représentés par plusieurs thèmes chacun et nous illustrons l'incorporation de cette multithématicité aux HMMs. La section 3 présente le modèle utilisé pour identifier de manière non-supervisée des sous-thèmes au sein des discours des deux présidents. Dans cette même section, nous analysons les problèmes d'estimation que pose ce modèle et proposons une solution originale pour y faire face. La section 4 contient une présentation complète du système utilisé pour la tâche DEFT et des performances qu'il nous a permis d'obtenir. Une analyse plus détaillée des résultats, permettant d'apprécier la contribution des différents outils utilisés aux résultats est également proposée. Cet article se conclut par une discussion des voies d'amélioration du système, dans la perspective d'autres tâches de fouille de texte.

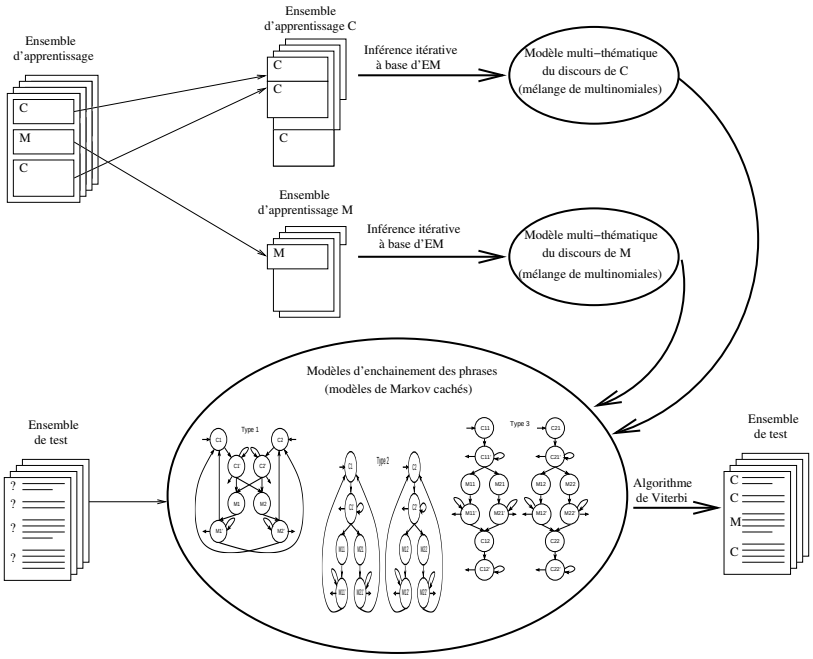

FIG. 1 – *Schéma global de l'approche*

2 Modèles de Markov pour la segmentation

L'approche que nous avons retenue combine deux outils de base de la fouille de textes, à savoir d'une part les modèles de Markov pour les séquences ; d'autre part les modèles de classification probabilistes non-supervisés. Dans cette section, nous discutons la conception et la mise en œuvre de modèles de Markov dont la topologie est spécifiquement adaptée à la tâche de segmentation DEFT. Partant d'un modèle très simple envisageant la segmentation sous l'angle de la catégorisation, nous introduisons progressivement des modèles plus complexes, qui prennent en compte les différentes contraintes de la tâche.

2.1 Un modèle simpliste de catégorisation

Le modèle le plus simple implantant les principes que nous avons retenus considère chaque pseudo-document comme une succession ordonnée de vecteurs C_p multidimensionnels (un par phrase). Chaque vecteur C_p contient le nombre d'occurrences $C_p(w)$, dans la phrase p (de longueur l_p), de chacun des mots w d'un vocabulaire d'indexation prédéfini (de taille n_W). La tâche consiste alors à répartir ces vecteurs en deux classes : la classe C (les phrases attribués à J. Chirac) et la classe M (celles de F. Mitterrand).

Le classifieur dit « Bayésien Naïf » (Lewis, 1998; McCallum et Nigam, 1998) permet d'effectuer, phrase par phrase, cette affectation à partir du modèle probabiliste suivant. Il s'agit de déterminer la classe $y^* \in \{C, M\}$ qui est la plus probable compte tenu de l'observation

courante C_p, soit :

$$y^* = \underset{y}{\operatorname{argmax}}\, P(Y = y|C_p) = \underset{y}{\operatorname{argmax}}\, P(C_p|Y = y)P(Y = y) \quad (1)$$

Chaque vecteur est considéré comme la réalisation d'un tirage d'une loi multinomiale. Si les valeurs des paramètres $\{\beta_{wy}, w = 1 \ldots n_W, y \in \{C, M\}\}$, pour chacune des lois associées à ces deux classes, sont supposées connues, il vient :

$$P(C_p|Y = y) = \frac{l_p!}{\prod_{w=1}^{n_W} C_p(w)!} \beta_{wy}^{C_p(w)}$$

Cette formule exprime simplement le fait que chaque occurrence d'un mot w dans la phrase C_p contribue à la probabilité globale par un facteur β_{wy}. Intuitivement, plus ce facteur est grand, plus le mot w est "important" pour la classe y. Le rapport de factorielles est le facteur de normalisation classique des lois multinomiales (qui ne joue aucun rôle dans la classification car il ne dépend pas des paramètres de classes).

Si la valeur des probabilités *a priori* $P(Y = y), y \in \{C, M\}$, pour chacune des deux classes est également connue, il devient possible d'utiliser (1) pour ventiler les phrases.

Ce classifieur fournit exactement la même segmentation que la mise en œuvre du décodage par l'algorithme de Viterbi dans le Modèle de Markov représenté à la figure 2, sous l'hypothèse que :

- la loi d'émission associée à chaque état est $P(C_p|Y = y)$, calculée selon les principes exposés ci-dessus.
- les probabilités de transition entre états, ainsi que les probabilités initiales et finales ne dépendent que des probabilités a priori $P(Y = y)$[1].

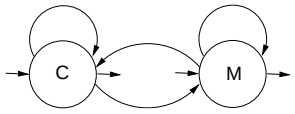

FIG. 2 – *HMM de base*
Les états ayant une transition entrante depuis les états initiaux sont identifiés par une flèche entrante ; ceux qui ont une transition vers l'état final ont une flèche sortante.

Nous allons voir dans ce qui suit qu'il est possible de préserver ce lien entre classifieurs et HMMs pour l'enchaînement des phrases tout en ajoutant de la complexité dans les deux étapes, avec l'ajout de sous-thèmes dans la classification (Section 3) d'une part et avec l'intégration de contraintes dans les transitions entre états (Sous-sections 2.2 et 2.3) d'autre part.

2.2 Intégration des contraintes de la tâche

L'approche décrite ci-dessus ignore entièrement le caractère séquentiel de la tâche, en posant l'hypothèse que chaque phrase est statistiquement indépendante des autres phrases du

[1] Ceci est vrai à un détail technique près, qui est que les probabilités finales sont liées de façon plus ou moins directe à la longueur du texte. Mais cette dépendance est accessoire dans l'analogie que nous établissons ici.

même pseudo-document. Cette hypothèse est très inappropriée, dans la mesure où les documents s'organisent comme une succession de passages relativement longs d'un même auteur. En jouant sur les probabilités de transition entre états de manière à (i) favoriser une des deux classes et (ii) pénaliser les changements de classes, il est aisé d'améliorer un peu le modèle de la figure 2 pour aboutir à un segmenteur fournissant un étiquetage plus cohérent temporellement.

Cette approche ne permet toutefois pas de respecter différentes contraintes qui constituent pourtant des informations intéressantes sur la tâche :

1. Un texte commence toujours par une phrase de la classe C.
2. Chaque fragment contient toujours au moins deux phrases de la même classe (pas de phrase de F. Mitterrand isolée et au moins deux phrases de J. Chirac en début de texte).
3. Chaque pseudo-document contient au plus une insertion d'un bloc de phrases de la classe M[2].

Il est possible d'intégrer progressivement ces contraintes en conservant l'architecture générale du système. La première s'exprime simplement par le fait que la probabilité initiale de l'état C est 1. La seconde se modélise en dupliquant chaque état (et les lois d'émission associées) pour donner lieu au modèle de la figure 3.(a), dans lequel les états C et C' (respectivement M et M') sont des clones de l'état C (respectivement M) de la figure 2. Dans ce nouveau modèle, tout fragment est ainsi contraint à « consommer » au moins deux phrases. La troisième contrainte est rendue explicite dans le modèle représenté Figure 3.(b), qui contient maintenant 4 clones de l'état C : $C1$ et $C1'$, qui modélisent les phrases de la classe C pré-insertion, et $C2$ et $C2'$ pour les phrases post-insertion.

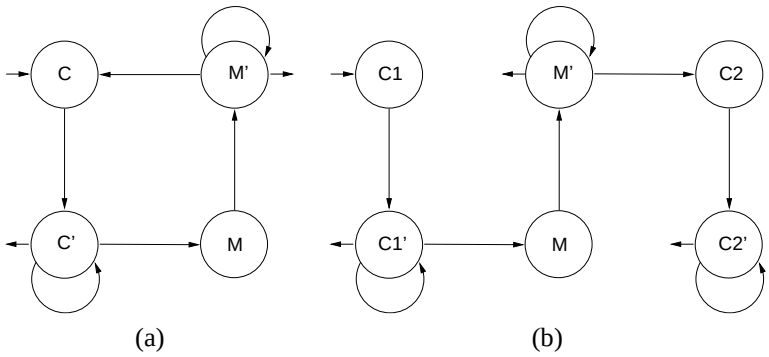

FIG. 3 – *HMMs avec contraintes*

2.3 Subdivision des classes C et M

Un examen attentif de la méthode de constitution des corpus d'apprentissage et de test révèle une autre information, qui est que l'ensemble des discours se ventile globalement en deux grands thèmes : 'national' et 'international' et que la procédure de constitution des corpus

[2]Cette contrainte ne figure pas explicitement dans la description de la tâche DEFT. Il nous a semblé intéressant de l'inclure, dans la mesure où tous les documents du corpus d'apprentissage la respectent.

a veillé à ce que chaque fragment ajouté (classe M) diffère thématiquement du discours dans lequel il s'insère : soit un fragment M-'national' dans un discours C-'international', soit le contraire. Cette autre clé de répartition des discours n'est malheureusement pas fournie par les organisateurs : pour utiliser cette information, il faudra donc, d'une certaine manière, la reconstruire de façon non-supervisée. Nous verrons précisément comment cela est possible à la section 3.

Supposons pour l'instant que cette information soit disponible, modélisée par quatre (et non plus deux) classes différentes : CN, CI, MN, et MI, à chacune desquelles est associé un modèle multinomial différent. La segmentation du texte est alors opérée par le modèle de la figure 4. En intégrant les trois contraintes décrites dans la section précédente, nous aboutirions à un modèle dont la topologie contiendra deux copies du modèle 3.(b), l'une pour les séquences $CN - MI - CN$, l'autre pour les séquences $CI - MN - CI$.

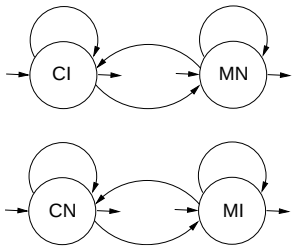

FIG. 4 – *HMMs avec contraintes et thèmes*

Dans la mesure où la répartition des documents en thèmes est inconnue, les modèles que nous avons finalement utilisés sont moins contraints. Ils reposent en effet sur l'hypothèse que les discours de J. Chirac se répartissent en n_C thèmes, et ceux de F. Mitterrand en n_M thèmes, l'observation selon laquelle 'national' et 'international' sont alternés se traduisant plus généralement par :

4. Toute transition directe entre deux thèmes du même auteur est interdite (à l'exception des insertions, chaque texte est donc supposé monothématique).
5. L'insertion d'un fragment de F. Mitterrand sépare deux fragments de discours de J. Chirac qui appartiennent au même thème.

Ces contraintes s'ajoutent à celles que nous avons énoncées dans la section précédente.

Pour valider l'apport en termes de performances de ces différentes hypothèses, nous avons finalement développé et testé 3 modèles, appelés "types" 1, 2 et 3 dans ce qui suit. Ils respectent tous les conditions 1, 2 et 4 mais :

- le type 1 ne tient pas compte des contraintes 3 (au plus une insertion M) et 5 (cohérence sous-thématique) ;
- le type 2 respecte la contrainte 5 mais pas 3 ;
- le type 3 respecte l'ensemble des contraintes discutées ci-dessus.

Ces modèles sont représentés sur les figures 5 et 6 dans le cas simple où $n_M = n_C = 2$.

La mise en œuvre de ces modèles requiert deux types de ressources : les paramètres des lois d'émission multinomiales, dont l'estimation est discutée à la section 3, et les probabilités de transition du HMM, qui sont calculées comme suit. Pour les probabilités de transition, nous

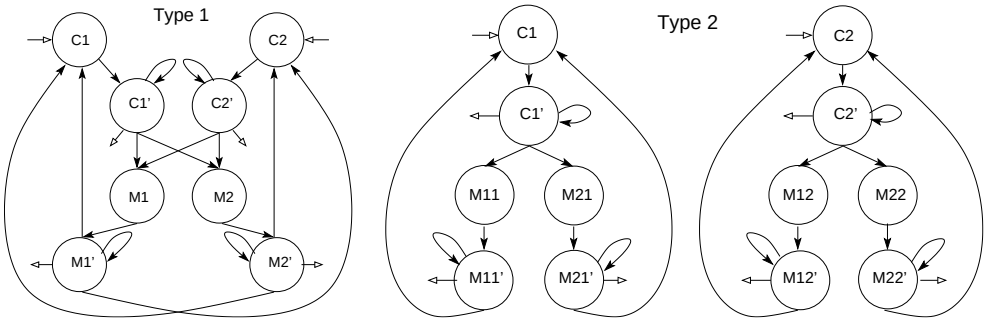

FIG. 5 – *Contraintes de "mono-thématicité" et de "retour" dans un même thème*
Modèles incluant les contraintes de non transition d'un thème à un autre pour un passage donné d'un interlocuteur (1) et d'identité des thèmes pré et post-insertion (2).

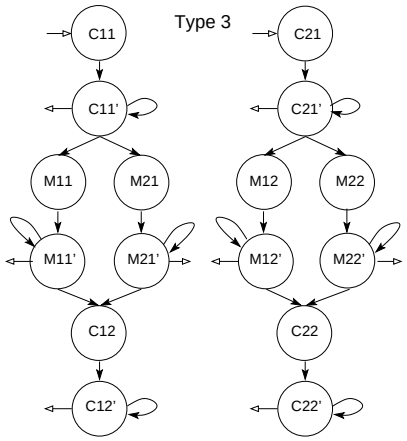

FIG. 6 – *Contrainte d'insertion unique*
Modèle respectant, en plus de toutes les autres, la contrainte d'insertion unique.

multiplierons les constantes de changement d'auteur (p_{C2M} et p_{M2C}, fixées à 0.3) par le paramètre α_t correspondant au thème vers lequel la transition s'effectue[3]. Les probabilités de sortie ont été en général fixées à 0, à l'exception notable du type 3 où les probabilités de sortie des états C post-insertion doivent être augmentées et fixées à p_{C2M}. En effet, dans le cas contraire, on favorise les états post-insertion par rapport aux états pré-insertion et la vraisemblance est alors presque toujours maximisée en affectant les deux premières phrases à J. Chirac, les deux suivantes à F. Mitterrand et toutes les autres à J. Chirac (seule configuration admissible qui maximise le nombre de paragraphes dans les états post-insertion). Les probabilités de rester

[3]Les α_t correspondent aux poids du mélange ; le problème de leur estimation sera également traité en section 3. Pour le moment, précisons simplement qu'il existe un α_t par sous-thème, donc $n_M + n_C$ au total, et que les n_M paramètres correspondant aux sous-thèmes M somment à 1, ainsi que les n_C paramètres correspondant aux sous-thèmes C.

dans le même état (boucle) sont calculées pour que la somme des probabilités de transition pour chaque état soit égale à 1.

3 Modèle de mélange de multinomiales

Nous exposons et discutons à présent la méthodologie utilisée pour construire, de manière non-supervisée, des sous-thèmes caractérisés par des distributions multinomiales à partir d'une collection de documents. Nous utilisons dans cette section de façon générique le terme document, mais il se peut très bien que chaque vecteur de comptes représente un paragraphe ou une phrase. Après une rapide présentation du modèle, nous discutons plus longuement des problèmes d'estimation et des solutions mises en œuvre.

Aux côtés des méthodes classiques de classification non-supervisée, telles que l'algorithme des K-moyennes ou l'Analyse en Composantes Principales (ou une variante proche : l'Analyse Sémantique Latente (Deerwester et al., 1990)), des méthodes probabilistes ont trouvé leur place pour l'analyse exploratoire de données textuelles, les modèles les plus populaires étant probablement *Probabilistic Latent Semantic Analysis* (Hofmann, 2001) et *Latent Dirichlet Allocation* (Blei et al., 2002). À l'instar de ces auteurs, nous nous plaçons ici dans le domaine des statistiques paramétriques et utilisons un modèle de mélange dont les variables latentes ont une signification *thématique*. Les paramètres de ces modèles ont ainsi une interprétation simple et l'on peut associer à chaque thème une distribution sur le vocabulaire qui identifie les mots les plus représentatifs pour ce thème. De manière duale, le résultat du partitionnement est, pour chaque document, un vecteur probabilisé d'appartenance aux différents thèmes, qui peut s'interpréter comme une projection du document dans un espace de faible dimension.

Nous considérons ici le plus simple de ces modèles probabilistes (Nigam et al., 2000; Clérot et al., 2004), dans lequel chaque document est supposé monothématique : ce modèle est ainsi directement compatible avec la tâche DEFT, dans lequel chaque phrase est supposée appartenir à un thème unique. Après avoir présenté ce modèle, nous donnons les équations d'estimation du Maximum A Posteriori, via l'algorithme Expectation Maximization (EM). Nous évoquons ensuite quelques résultats qui montrent l'importance de l'initialisation et suggérons une méthode heuristique pour l'inférence des paramètres, qui est celle que nous avons finalement utilisée pour les évaluations.

3.1 Le modèle génératif

Nous supposons toujours que les textes sont représentés par des « sacs-de-mots », c'est-à-dire que le vocabulaire est connu et fini et que chaque document est représenté par un vecteur de comptes sur cet ensemble. Conformément aux notations introduites en 2.1, n_D, n_T et n_W représentent respectivement le nombre de documents dans le corpus d'apprentissage, le nombre de thèmes (i.e. le nombre de composantes du modèle de mélange) et la taille du vocabulaire.

Pour $d \in \{1, \ldots, n_D\}$ et $w \in \{1, \ldots, n_W\}$, on note $C_d(w)$ le terme général de la matrice de comptes, c'est-à-dire le nombre d'occurrences du mot w dans le document d'indice d. Désignons également par $l_d = \sum_{w=1}^{n_W} C_d(w)$ le nombre d'occurrences dans le texte d et par $l = \sum_{d=1}^{n_D} l_d$ le nombre total d'occurrences dans le corpus, somme de tous les termes de la matrice de comptes.

Contrairement au cadre de la catégorisation supervisée, une hypothèse faite ici est qu'aucune information sur les affectations des documents aux différents thèmes n'est à notre disposition. Le modèle de génération du corpus présenté ci-dessous permet, après estimation des paramètres, de proposer une ventilation des documents suivant les différentes composantes du mélange.

On suppose que les textes sont indépendants. Le document numéroté d ($d \in \{1, \ldots, n_D\}$) résulte de l_d tirages indépendants sur le vocabulaire selon une distribution dépendant du thème, ce dernier étant défini par une variable cachée tirée une fois par texte. Notons $\mathrm{Mult}(k, (p_1, \ldots, p_n))$ l'opération consistant à tirer k fois suivant une multinomiale de probabilités $(p_1, \ldots, p_n)$. D'où le modèle génératif pour un document :

- Tirer un thème $T \sim \mathrm{Mult}(1, (\alpha_1, \ldots, \alpha_{n_T}))$ où les α_t sont des paramètres tels que $\sum_{t=1}^{n_T} \alpha_t = 1$.
- Tirer l_d mots $C_d = (C_{d1}, \ldots, C_{dn_W}) \sim \mathrm{Mult}(l_d, (\beta_{1t}, \ldots, \beta_{n_W t}))$, β étant une matrice $n_W \times n_T$ de paramètres telle que $\forall t \in \{1, \ldots, n_T\}, \sum_{w=1}^{n_W} \beta_{wt} = 1$.

La probabilité d'un document est alors, en notant T_d la variable indicatrice du thème latent :

$$\begin{aligned} P(C_d; \alpha, \beta) &= \sum_{t=1}^{n_T} P(T_d = t; \alpha, \beta) P(C_{d1}, \ldots, C_{dn_W} | T_d = t; \alpha, \beta) \\ &= \sum_{t=1}^{n_T} P(T_d = t; \alpha, \beta) l_d! \prod_{w=1}^{n_W} \frac{P(w|T_d = t; \alpha, \beta)^{C_d(w)}}{C_d(w)!} \\ &= \frac{l_d!}{\prod_{w=1}^{n_W} C_d(w)!} \sum_{t=1}^{n_T} \alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)} \end{aligned}$$

Ainsi chaque thème contribue à la probabilité globale du document par sa probabilité a priori α_t et, pour chaque occurrence du texte, par la probabilité β_{wt} d'émission du mot w dans le thème en question.

La probabilité du corpus, ou vraisemblance des observations, est obtenue en réalisant le produit de l'expression ci-dessus pour l'ensemble des documents étudiés. Cependant, il n'est pas possible d'établir directement une expression d'un estimateur de maximum de vraisemblance. C'est pourquoi nous faisons appel à l'algorithme EM (Expectation Maximization) qui repose sur le calcul de l'espérance, conditionnellement aux observations, de la log-vraisemblance *complète* $\mathcal{L}^c$, c'est-à-dire la log-vraisemblance des couples vecteur de comptes C_d et thème T_d, définie par :

$$\begin{aligned} \mathcal{L}^c &= \sum_{d=1}^{n_D} \log P(C_d, T_d) \\ &= \sum_{d=1}^{n_D} \Big(\log P(T_d) + \log P(C_d|T_d) \Big) \\ &= \sum_{d=1}^{n_D} \sum_{t=1}^{n_T} \mathbb{1}_{\{T_d = t\}} \left(\log \alpha_t + \sum_{w=1}^{n_W} C_d(w) \log \beta_{wt} \right) + K \end{aligned}$$

où K est une constante indépendante des paramètres (que nous oublierons par la suite). La notation $\mathbb{1}_A$ désigne la fonction indicatrice définie par :

$$\mathbb{1}_A = \begin{cases} 1 & \text{si } A \text{ est vrai ;} \\ 0 & \text{sinon.} \end{cases}$$

L'espérance, conditionnellement aux observations, et tenant compte des paramètres α', β' issus de l'itération précédente, s'écrit :

$$E[\mathcal{L}^c] = \sum_{d=1}^{n_D} \sum_{t=1}^{n_T} P(T_d = t | C_d; \alpha', \beta') \times \left(\log \alpha_t + \sum_{w=1}^{n_W} C_d(w) \log \beta_{wt} \right)$$

Les probabilités *a posteriori* sont données par la formule de Bayes, conduisant, pour $t \in \{1, \ldots, n_T\}, d \in \{1, \ldots, n_D\}$, à :

$$\begin{aligned} P(T_d = t | C_d; \alpha', \beta') &= \frac{P(C_d | T_d = t; \alpha', \beta') P(T_d = t; \alpha', \beta')}{P(C_d; \alpha', \beta')} \\ &= \frac{P(C_d | T_d = t; \alpha', \beta') P(T_d = t; \alpha', \beta')}{\sum_{t'=1}^{n_T} P(C_d | T_d = t'; \alpha', \beta') P(T_d = t'; \alpha', \beta')} \\ &= \frac{\alpha'_t \prod_{w=1}^{n_W} {\beta'_{wt}}^{C_d(w)}}{\sum_{t'=1}^{n_T} \alpha'_{t'} \prod_{w=1}^{n_W} {\beta'_{wt'}}^{C_d(w)}} \end{aligned} \quad (2)$$

Le numérateur, produit de la probabilité a priori du thème t et de l'"importance" dans le thème de chaque occurrence du mot w dans le texte considéré, a déjà été identifié précédemment comme mesurant intuitivement la probabilité jointe du document C_d et du thème t. Le dénominateur vient de l'opération de normalisation correspondant à $\sum_{t=1}^{n_T} P(T_d = t | C_d; \alpha', \beta')$.

Il est alors possible de déterminer les équations de ré-estimation des paramètres en maximisant la quantité de l'EM, avec la technique des multiplicateurs de Lagrange pour normaliser de façon appropriée les paramètres α (le vecteur somme à 1) et β (chaque colonne somme à 1). On obtient, pour $t \in \{1, \ldots, n_T\}$ et $w \in \{1, \ldots, n_W\}$:

$$\alpha_t = \frac{1}{n_D} \sum_{d=1}^{n_D} P(T_d = t | C_d; \alpha', \beta') \quad (3)$$

$$\beta_{wt} = \frac{\sum_{d=1}^{n_D} C_d(w) P(T_d = t | C_d; \alpha', \beta')}{\sum_{w=1}^{n_W} \sum_{d=1}^{n_D} C_d(w) P(T_d = t | C_d; \alpha', \beta')} \quad (4)$$

Ces formules ont elles aussi une interprétation intuitive simple si les probabilités d'appartenance $P(T_d = t | C_d; \alpha', \beta')$ sont exactement 0 ou 1 (chaque texte "appartient" alors à un thème et un seul) :

- On obtient α_t en comptant le nombre de documents dans le thème t puis en normalisant.
- On détermine la nouvelle valeur de β_{wt} en dénombrant le nombre d'occurrence du mot w dans les textes correspondant au thème t, puis on normalise sur l'ensemble des mots.

Cette interprétation peut être étendue au cas où les probabilités d'appartenance ne sont pas binaires. Chaque texte contribue alors au renouvellement des paramètres en proportion de son "implication" dans le thème.

L'algorithme EM consiste à appliquer les formules (2), (3) et (4) de façon itérative jusqu'à convergence.

Lorsqu'un mot w n'est jamais observé dans un thème t, ces formules conduisent à une estimation nulle pour β_{wt}. Il est alors nécessaire de recourir à des techniques de lissage des estimateurs, pour rendre compte du fait que, même si, dans l'ensemble d'apprentissage, un mot n'a jamais été vu en conjonction avec un thème donné, son apparition dans ce thème n'est pas totalement impossible (elle est néanmoins de probabilité très faible). Dans la suite, nous utilisons un lissage de Laplace, consistant à augmenter tous les comptes de 0.1. Dans le cadre probabiliste, ce lissage peut être interprété comme correspondant à l'algorithme EM associé au maximum *a posteriori* (et non plus au maximum de vraisemblance) lorsque les paramètres β sont munis d'une distribution *a priori* de type Dirichlet, de paramètre 1.1 (Rigouste et al., 2005a).

3.2 Méthode d'inférence itérative par ajout de mots rares

Les équations de ré-estimation posées, il reste encore une marge de manœuvre importante pour un expérimentateur désirant inférer les paramètres du modèle. Des questions pertinentes concernent notamment :

- le choix du vocabulaire : faut-il considérer le vocabulaire en entier ou retirer les mots trop rares ou trop fréquents ?
- l'initialisation du modèle

Ces interrogations sont étudiées en détail dans (Rigouste et al., 2005b) : nous résumons dans un premier temps les conclusions de ces travaux qui nous semblent pertinentes pour DEFT, avant de présenter la méthode d'inférence finalement utilisée pour les tests.

3.2.1 Expérimentations préliminaires

Le corpus qui a servi de base à ces expérimentations préliminaires est un corpus raisonnablement simple, issu de Reuters 2000[4] et composé de 5000 textes équirépartis dans 5 catégories (arts, sports, emploi, catastrophes, santé) (Reuters, 2000). En plus de la log-vraisemblance (ou de manière équivalente, de la perplexité[5]) calculée lors de l'apprentissage, nous considérons également une autre mesure des performances : *l'information mutuelle* entre le classement produit par le modèle et les catégories du corpus Reuters. Ce critère mesure plus directement la faculté de l'algorithme à retrouver les regroupements d'origine.

Nos expériences nous ont permis de mettre en évidence le fait que la phase d'initialisation de l'algorithme EM est cruciale pour obtenir des regroupements pertinents des documents. Elles ont également confirmé l'intuition suivante : en l'absence d'information *a priori* sur les thèmes à trouver, la meilleure initialisation consiste à partir de regroupements qui se recoupent largement, l'apprentissage se chargeant en général de les séparer. Dans cet esprit, l'algorithme est initialisé en fixant les probabilités *a posteriori* pour un document d'appartenir à un thème – équation (2) – très proches de l'équiprobabilité entre tous les thèmes. Pour chaque essai, nous tirons donc ces valeurs selon une distribution Dirichlet dont la moyenne est la distribution

[4]Le corpus est en anglais. Savoir si les conclusions de notre étude se transposent à un autre corpus, en français, comme nous l'avons supposé, reste une question ouverte.

[5]La perplexité est définie comme l'exponentielle de la valeur moyenne (par mot) de la log-vraisemblance (Jelinek, 1997).

uniforme et dont la variance est faible. C'est-à-dire que si l'on fixe un paramètre $\lambda > 0$ grand, les probabilités a posteriori sont tirées pour chaque document selon une densité g telle que :

$$g(p_1, \ldots, p_{n_T}) \propto \prod_{t=1}^{n_T} {p_t}^{\lambda - 1}$$

Ainsi, plus λ est grand, plus on obtient des probabilités proches de l'équiprobabilité : on peut en effet montrer que la moyenne de chaque p_t est $\frac{1}{n_T}$ et que sa variance est équivalente à $\frac{1}{\lambda}$ lorsque $\lambda \gg n_T$. Par la suite, nous désignerons ce procédé sous le nom d'initialisation "Dirichlet".

Afin d'avoir une idée de la meilleure performance possible, nous avons également essayé d'introduire l'information de supervision disponible, consistant à baser l'initialisation sur les catégories Reuters. Pour ce faire, l'étape d'initialisation donne à un document d de catégorie Reuters t une valeur de 1 à la probabilité *a posteriori* d'appartenir au thème t, et une valeur 0 pour tous les autres thèmes.

La comparaison systématique de ces différentes procédures d'initialisation a conduit à établir les constats suivants :

- la variabilité entre les deux initialisations est très forte pour les deux mesures : log-vraisemblance et information mutuelle.
- la log-vraisemblance mesurée sur l'ensemble d'apprentissage est un indicateur raisonnable de la qualité finale du regroupement produit, dans le sens où ses variations sont comparables à celles de la log-vraisemblance sur les données de test ou de l'information mutuelle sur les données d'apprentissage ou de test.
- à moins de pouvoir les initialiser correctement (ce qui est impossible sans information de supervision), les mots rares nuisent en général à l'apprentissage et l'écart entre les deux initialisations diminue lorsque l'on réduit la taille du vocabulaire en ne conservant que les mots les plus fréquents.

3.2.2 Une nouvelle stratégie d'inférence

Sur la base de ces observations, la méthode d'inférence finalement retenue repose sur le principe d'une augmentation progressive de la taille du vocabulaire d'indexation (une présentation plus complète est donnée dans (Rigouste et al., 2005b)). Son principe est le suivant : partant d'un vocabulaire extrêmement réduit (constitué des 1000 mots les plus fréquents, soit 2% du vocabulaire total), une première estimation des paramètres du modèle est obtenue avec l'initialisation "Dirichlet". Ce procédé est répété plusieurs fois et seul le meilleur ensemble de paramètres (au sens de la log-vraisemblance finale) est conservé. Au terme de cette étape, nous disposons donc d'une valeur pour un (petit) sous-ensemble des paramètres β_{wt} correspondant aux mots les plus fréquents. Il est possible d'en déduire, par application de l'étape M de l'algorithme EM (équation (2)), des probabilités a posteriori d'appartenance aux thèmes pour *tous les documents*. Après augmentation de la taille du vocabulaire, nous prenons alors soin d'utiliser une initialisation "Dirichlet" dont l'espérance n'est plus maintenant uniforme, mais égale aux probabilités a posteriori de l'étape précédente. L'algorithme EM peut alors être mis en œuvre pour ré-estimer les paramètres β_{wt} pour un plus grand ensemble de mots, desquels seront déduites de nouvelles probabilités a posteriori, etc. Cette procédure est itérée jusqu'à ce que le vocabulaire complet soit finalement pris en compte.

Les résultats présentés dans (Rigouste et al., 2005b) montrent que l'algorithme d'inférence itératif parvient à atteindre les mêmes valeurs de vraisemblance que celles obtenues en initialisant avec les informations de supervision. L'information mutuelle est un peu moins bonne, montrant que la corrélation entre les deux indicateurs n'est pas absolue, mais se situe dans des plages de valeurs beaucoup plus satisfaisantes qu'avec l'initialisation "Dirichlet" simple.

4 Utilisation du segmenteur en thèmes pour DEFT

L'idée directrice de notre méthode est qu'il devrait être plus facile d'identifier les ruptures thématiques entre les phrases prononcées par J. Chirac et celles de F. Mitterrand si l'on connaît précisément les différents sujets abordés par chaque locuteur. Nous pensons (et cela se confirme dans la dernière section) que le résultat sera meilleur en modélisant les discours de chaque président par plusieurs thèmes, qui lui sont propres, plutôt qu'en utilisant seulement un thème pour chaque personne.

Ainsi, nous utilisons les données d'apprentissage pour estimer les paramètres relatifs aux thèmes abordés par J. Chirac et à ceux abordés par F. Mitterrand. Une fois ces paramètres identifiés, nous utilisons l'algorithme de Viterbi sur les phrases du corpus de test pour déterminer le thème (et donc l'auteur) le plus vraisemblable pour chaque phrase.

4.1 Prétraitements

Rappelons que la campagne DEFT inclut trois tâches, soit, par ordre de difficulté croissante : la tâche 3 consiste à segmenter des textes "bruts" ; la tâche 2 des textes dans lesquels les dates sont remplacées par un tag unique <date> ; la tâche 1 des textes dans lesquels ont également été normalisés et remplacés par le tag <nom> les noms de personnes. Pour chacune des trois tâches, la même série de prétraitements des corpus a été utilisée, consistant à segmenter chaque phrase en mots, à normaliser les chiffres, à mettre tous les mots en minuscules et à supprimer toutes les marques de ponctuation.

À l'issue de ces traitements, le vocabulaire utilisé dans les modèles statistiques peut être identifié : il contient toutes les formes qui apparaissent dans le corpus, y compris les mots-outils et les mots rares, soit environ 30 000 formes graphiques. Lorsqu'un document du corpus de test contient un mot qui n'apparaît pas dans le corpus d'entraînement, ce mot est simplement ignoré.

Nous formulons l'hypothèse que, dans un fichier de l'ensemble d'entraînement, toutes les phrases prononcées par un président donné font partie du même thème. Par conséquent, le corpus d'entraînement pour apprendre les sous-thèmes de J. Chirac est constitué en supprimant les insertions de F. Mitterrand et en agrégeant les parties de texte séparées par ces insertions. Deux passages qui appartiennent à deux documents différents dans le corpus original ne sont jamais concaténés dans le même texte. De la même manière, chaque fragment attribué à F. Mitterrand constitue un document distinct.

4.2 Description de l'algorithme

L'algorithme itératif d'estimation des paramètres décrit en section 3.2 est utilisé pour obtenir les coefficients β_{wt} correspondant aux thèmes récurrents des discours de J. Chirac. Le

nombre de thèmes n_C est fixé *a priori* en effectuant des mesures de F-score (tel que défini par les organisateurs pour l'évaluation de la tâche DEFT) en validation croisée sur l'ensemble d'apprentissage. Nous déterminons de même un nombre de thèmes n_M pour F. Mitterrand, ce qui donne au total $n_C + n_M$ distributions sur le vocabulaire, qui sont représentatives des différents sujets abordés dans les discours de J. Chirac et F. Mitterrand.

Sur les textes du corpus de test, nous n'avons d'autre choix que d'affecter une variable latente à chaque phrase puisque nous n'avons pas d'information a priori sur les ruptures thématiques. Le problème est alors d'évaluer la séquence thématique la plus probable pour chaque nouveau texte. Pour cela, nous utilisons les modèles de Markov cachés qui ont été présentés à la section 2, en utilisant le modèle de mélange de la section 3 pour calculer la vraisemblance des phrases examinées dans chaque thème. Pour chaque document du corpus de test, l'état (c'est-à-dire le thème) le plus probable de chaque phrase est ainsi déterminé par application de l'algorithme de Viterbi.

4.3 Évaluation du segmenteur

Nous étudions ici uniquement les résultats sur la tâche 1 de DEFT : dans la mesure où nous n'avons pas cherché à tirer profit des informations spécifiques liées aux noms et aux dates, nos performances sur les autres tâches sont quasiment identiques à celles obtenues sur la tâche 1. Sur la base des résultats du Défi (Alphonse et al., 2005), il semble que, comparativement aux autres méthodes, notre modèle soit plus efficace sur cette tâche que sur les deux autres, dans la mesure où nos scores sont sensiblement les mêmes sur les trois tâches alors que d'autres équipes améliorent leurs performances sur les tâches 2 et 3. Pour faire de même, il nous aurait fallu ajuster les poids des dates et des noms de personnes dans le calcul de la vraisemblance de chaque phrase.

Pour la campagne de test officielle, nous avons soumis sur cette tâche les types 1 et 2 avec $n_C = 10$ et $n_M = 4$, le type 2 obtenant les meilleures performances. La table 1 montre qu'il est possible d'atteindre des performances légèrement supérieures en utilisant le type 3, toujours pour les mêmes nombres de thèmes.

	F	Préc.	Rappel	Corr.	M → C	C → M
Type 1 (soumission DEFT 1)	80.33	88.01	73.87	94.94	3.65	1.41
Type 2 (soumission DEFT 2)	86.04	86.44	85.65	96.12	2.01	1.88
Type 3	86.96	84.32	89.78	96.24	1.43	2.33
Vainqueurs DEFT	87.0	88.3	85.8	–	–	–

TAB. 1 – *Évaluation des différents modèles*
Résultats pour $n_C = 10$ et $n_M = 4$ et comparaison avec la meilleure équipe. En plus des valeurs de F-score (pour $\beta = 1$), pour chaque modèle sont donnés la précision, le rappel (également tels que définis par les organisateurs de DEFT), ainsi que le pourcentage de phrases correctes et les pourcentages d'erreurs par type (phrases de J. Chirac attribuées à F. Mitterrand, et phrases de F. Mitterrand attribuées à J. Chirac).

Notons que le meilleur résultat parmi tous nos essais, autour de $F = 88$ (non présenté dans la table 1), est obtenu avec le type 3 en fixant $n_C = 10$ et $n_M = 3$.

De façon générale, les résultats de DEFT (Alphonse et al., 2005) valident largement l'approche consistant à combiner méthodes de classification et de segmentation. Plus spécifiquement, on remarque que les équipes utilisant des HMMs (El-Bèze et al., 2005; Labadié et al., 2005; Kerloch et Gallinari, 2005) obtiennent toutes des résultats satisfaisants.

Une analyse plus fine des résultats obtenus avec le type 3 permet de constater que les erreurs ne sont pas également réparties, mais se concentrent sur un nombre relativement faible de documents. Elles correspondent à des situations dans lesquelles le modèle a choisi de considérer qu'un fragment significatif du texte était une insertion, alors que le texte n'en comprenait pas ; ou bien, au contraire, a failli à détecter la moindre insertion. Près de 90% des erreurs sont ainsi localisées dans les 80 documents de test les moins bien étiquetés, alors qu'à l'inverse plus de la moitié des documents ne contiennent aucune erreur de segmentation.

4.4 Analyse des résultats

La figure 7 permet d'apprécier quantitativement l'apport de chacune des contraintes ainsi que de l'augmentation du nombre de sous-thèmes. Il apparaît ainsi que multiplier les sous-thèmes est d'autant plus efficace que le modèle utilisé est contraint : pour le type 1, les résultats obtenus en prenant un thème par classe sont pratiquement les meilleurs ; alors que pour le type 3, nous observons une augmentation sensible des performances lorsque nous multiplions les sous-divisions thématiques. Dans tous les cas, la solution consistant à ne garder que deux sous-thèmes par classe est clairement sous-optimale. Il semble que les nombres optimaux de sous-thèmes soient de 3 pour la classe M et de 8 ou 10 pour la classe C. Cette différence s'explique probablement par la quantité de données d'apprentissage, bien plus importante pour un auteur que pour l'autre, et qui permet par conséquent d'estimer de façon fiable un plus grand nombre de paramètres.

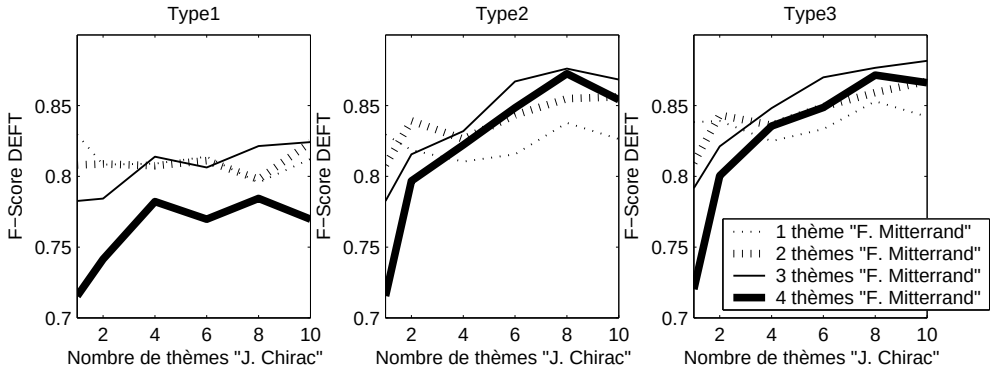

FIG. 7 – *F-Score obtenu sur la tâche 1 de DEFT en fonction de n_C*

En dehors de ces évaluations portant sur la tâche DEFT, une question légitime, et particulièrement pertinente en ce qui concerne l'utilité pratique de ce type de modèles, concerne la nature des thèmes déterminés de façon non supervisée lors de l'apprentissage. La procédure d'interprétation des thèmes que nous avons adoptée repose sur une heuristique en deux étapes :

– partant de la représentation complète des documents en sacs de mots, nous utilisons l'approche de sélection de variables décrite dans (Boullé, 2006), qui permet de ne retenir que les attributs (les mots) qui, individuellement, apportent une information vis-à-vis de la classification des documents sur les 14 thèmes. Cette étape vise à réduire fortement la taille du vocabulaire « discriminant » ;
– à partir de cette représentation réduite et pour chacun des thèmes, nous calculons la matrice de co-occurrences entre les mots des documents du thème, qui est interprétée comme une matrice de similarité relativement au thème. Pour caractériser le thème, nous retenons finalement les N mots les plus centraux au sens de la centralité spectrale, consistant à considérer la projection du mot sur la première direction propre de la matrice de similarité.

Les tables 2 et 3 listent les 15 mots les plus saillants ainsi déterminés pour chaque thème. On peut constater que, dans la plupart des cas, ces mots-clés suffisent à donner une idée assez précise de la thématique sous-jacente. C'est en particulier toujours le cas pour les discours de J. Chirac (thèmes C1 à C10) ; l'interprétation de certains thèmes des discours de F. Mitterrand est moins nette (voir le thème M2 en particulier). Mais, même dans ce dernier cas, le classement suit une bipartition entre thèmes plus nationaux (M1 et M2) et internationaux (M3 et M4).

Cette analyse permet de mettre en évidence que la méthode de classification employée, bien que n'exploitant aucune information de nature sémantique, permet effectivement de construire des groupes de documents présentant une unité thématique.

Enfin, pour évaluer l'apport de l'algorithme itératif d'initialisation utilisé pour apprendre les paramètres des thèmes, nous avons comparé ses performances avec celles obtenues en utilisant une procédure d'initialisation plus simple (initialisation « Dirichlet »), qui considère d'emblée tout le vocabulaire. Les mesures utilisées sont la perplexité (capacité de prédiction du modèle probabiliste) et l'information mutuelle (coïncidence des étiquettes C/M proposées avec les "vraies" classes). Comme le montrent les résultats de la table 4, moyennés sur 50 tirages, les performances nettement meilleures obtenues par la méthode itérative telles que les mesure la perplexité se traduisent également par un gain, d'ampleur plus modeste, sur la tâche d'évaluation extrinsèque de DEFT.

5 Conclusion

Nous avons présenté dans cet article la méthode utilisée pour répondre au problème posé dans le cadre du DÉfi Fouille de Textes 2005. Notre approche s'appuie sur l'utilisation de deux outils de base de la fouille de textes : d'une part les modèles de Markov cachés, d'autre part un modèle de classification non supervisée. En particulier, nous avons montré qu'en identifiant les distributions thématiques qui sous-tendent les discours de J. Chirac et F. Mitterrand dans le corpus d'entraînement, nous sommes mieux à même de segmenter le corpus de test, en déterminant l'enchaînement thématique le plus probable. Il est remarquable que ces deux modules n'aient pas été spécifiquement conçus pour ce travail, mais que leur assemblage permette d'aboutir à des résultats très satisfaisants.

Dans le cadre des tâches DEFT, de nombreuses améliorations de cette stratégie sont envisageables, consistant, par exemple, à apprendre de manière conjointe plutôt que séparée les différents paramètres (lois d'émission et probabilités de transition) des modèles de Markov.

Thème C1	Thème C2	Thème C3	Thème C4	Thème C5
amitié	école	cultures	mondialisation	états
peuples	familles	diversité	accord	union
paix	famille	côtés	amérique	conseil
continent	concitoyens	partenariat	croissance	européen
peuple	public	coopération	lutte	membres
histoire	égalité	paix	guerre	européenne
europe	citoyens	dialogue	marché	aide
deux	autorité	soutien	puissance	institutions
union	enfants	amis	peuples	internationale
ambition	nation	engagement	union	mise
relation	service	monsieur	indispensable	sommet
construire	publique	cher	règles	sécurité
vision	sociale	sommet	continent	défense
nations	services	projets	européen	européens
liberté	société	votre	emplois	mondiale
Thème C6	Thème C7	Thème C8	Thème C9	Thème C10
afrique	devoir	république	combat	moi
paris	protection	peuple	liberté	maire
moi	citoyens	cher	guerre	succès
partenariat	combat	emplois	général	veut
relation	autorité	compatriotes	fut	façon
choses	société	maire	valeurs	justice
guerre	public	mes	jamais	puis
fut	mission	amitié	homme	puissance
vivre	lutte	choses	paris	domaine
institutions	droit	veux	nom	lors
problèmes	sociale	europe	honneur	paix
membres	école	justice	forces	union
peuples	répondre	union	contre	démocratie
messieurs	famille	nation	justice	europe
mesdames	agit	construction	devoir	indispensable

TAB. 2 – *Mots caractéristiques pour les thèmes de J. Chirac*

Thème M1	Thème M2	Thème M3	Thème M4
sociaux	marché	europe	guerre
paix	on	cent	nations
renforcer	effort	communauté	états
coopération	cinq	puissance	peuples
sommet	entendu	marché	comment
économiques	simplement	trois	forces
devons	moins	serait	droit
niveau	plan	construction	équilibre
compatriotes	cent	institutions	peuple
hommage	quand	européen	droits
liens	faut	peu	paix
économie	plusieurs	était	amérique
notamment	mille	moins	sécurité
membres	là	après	cas
afrique	est-à-dire	autres	dès

TAB. 3 – *Mots caractéristiques pour les thèmes de F. Mitterrand*

Méthode	Perplexité - corpus C	Perplexité - corpus M	Information Mutuelle
Init. Dirichlet	755.7±3.4	775.5±2.2	0.83±0.01
Init. Itérative	733.3±2.2	760.8±2.2	0.85±0.01

TAB. 4 – *Résultats comparés pour deux méthodes d'inférence des paramètres*
Les mesures d'évaluation sont calculées sur un ensemble de test, puis moyennées sur plusieurs essais et en validation croisée.

Au-delà de ces aménagements immédiats, peu justifiés au vu du caractère un peu artificiel de la tâche proposée, nous envisageons d'explorer d'autres voies pour améliorer ces modèles.

Concernant les modèles séquentiels, nous prévoyons de tester sur d'autres tâches la combinaison de ce modèle de mélange thématique et d'un modèle de Markov caché modélisant l'enchaînement des variables latentes associées aux phrases ou aux paragraphes, dans un cadre dans lequel l'apprentissage de l'ensemble du modèle s'effectue de manière intégralement non-supervisée. Nous espérons ainsi montrer que les bons résultats obtenus ici, qui s'appuient en partie sur les spécificités structurelles du corpus (informations sur les règles d'insertion), sont généralisables à des applications moins contraintes. Concernant plus spécifiquement le modèle de classification probabiliste non-supervisée, nous envisageons de continuer à travailler sur l'amélioration des méthodes d'inférence, en comparant la méthode itérative heuristique présentée dans cet article avec des algorithmes de simulation (échantillonneur de Gibbs).

Remerciements Nous adressons nos remerciements aux organisateurs de ce Défi Fouille de Textes pour leur travail considérable, ainsi qu'aux rapporteurs dont les commentaires ont grandement contribué à l'amélioration de l'article.

Références

Alphonse, E., A. Amrani, J. Azé, T. Heitz, A.-D. Mezaour, et M. Roche (2005). Préparation des données et analyse des résultats de DEFT'05. In *"Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, Volume 2, pp. 99–111.

Benzécri et al., J.-P. (1981). *Pratique de l'analyse des données, tome 3. Linguistique et lexicologie*. Paris : Dunod.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2002). Latent Dirichlet allocation. In T. G. Dietterich, S. Becker, et Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS)*, Volume 14, Cambridge, MA, pp. 601–608. MIT Press.

Boullé, M. (to appear, 2006). MODL : a Bayes optimal discretization method for continuous attribute. *Machine Learning*.

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

Clérot, F., O. Collin, O. Cappé, et E. Moulines (2004). 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.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, et R. A. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41*(6), 391–407.

El-Bèze, M., J.-M. Torres-Moreno, et F. Béchet (2005). Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac. In *Dans les actes de la conférence "Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, Volume 2, pp. 125–134.

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

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

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press.

Kerloch, F. et P. Gallinari (2005). Extraction d'information à partir de modèles de Markov cachés. In *Dans les actes de la conférence "Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, Volume 2, pp. 145–153.

Labadié, A., Y. Romero, et L. Sitbon (2005). Segmentation et classification : deux politiques complémentaires. In *Dans les actes de la conférence "Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, Volume 2, pp. 183–192.

Lewis, D. D. (1998). Naive (Bayes) at forty : The independence assumption in Information Retrieval. In C. Nédellec et C. Rouveirol (Eds.), *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Number 1398, Chemnitz, DE, pp. 4–15. Springer Verlag, Heidelberg, DE.

Maisonnasse, L. et C. Tambellini (2005). Dépendances syntaxiques et méthodes de détection de passages pour une segmentation sur le locuteur et le thème. In *Dans les actes de la conférence "Traitement Automatique des Langues Naturelles" (TALN 2005) - Atelier DEFT'05*, Volume 2, pp. 155–164.

McCallum, A. et K. Nigam (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*.

Nigam, K., A. K. McCallum, S. Thrun, et T. M. Mitchell (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning 39*(2/3), 103–134.

Reuters (2000). Reuters corpus.

Rigouste, L., O. Cappé, et F. Yvon (2005a). 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.

Rigouste, L., O. Cappé, et F. Yvon (2005b). Inference for probabilistic unsupervised text clustering. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, France.

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

Summary

In this contribution, we show how we used generic probabilistic text mining tools to solve a supervised task: the DÉfi Fouille de Textes 2005. We first explain how the specificities of the task can be captured in the form of Hidden Markov Models. Then we present a probabilistic approach for text clustering, which models texts by a mixture of multinomial distributions over the word counts, where each component corresponds to a different theme. We apply the EM algorithm to estimate the parameters of these thematic distributions. This model is used to thematically subdivide the available training corpus in an unsupervised manner. We finally present and discuss the performance obtained using the combination of these tools.