Atelier DEFT'05, TALN 2005, Dourdan, 6-10 juin 2005

Modèle de mélange multi-thématique pour la Fouille de Textes

Loïs Rigouste, Olivier Cappé et François Yvon

Ecole Nationale Supérieure des Télécommunications
(GET /CNRS UMR 5141)
46 rue Barrault, 75634 Paris Cedex 13, France
(e-mails: `rigouste, cappe, yvon` at `enst.fr`)

Mots-clefs : modèle de mélange, distributions multinomiales, algorithme de Viterbi

Keywords: mixture model, multinomial distributions, Viterbi algorithm

Résumé Dans cet article, nous montrons comment des techniques probabilistes d'analyse exploratoire peuvent être utilisées pour résoudre une tâche d'apprentissage supervisée: le DÉfi Fouille de Textes 2005. Ainsi, nous présentons une modélisation des textes par un mélange de distributions multinomiales sur les comptes de mots, chaque composante correspondant à un thème particulier. Les paramètres des distributions thématiques sont estimés grâce à l'algorithme EM. Les thèmes étant appris séparement pour chacun des deux auteurs, un produit dérivé de l'identification des thèmes est l'attribution des documents à leur auteur putatif. Dans la phase de test, nous affectons à chaque phrase une variable latente, en prenant soin de lier les thèmes à l'intérieur d'un document par un modèle de Markov caché, dont les paramètres sont fixés a priori. La détermination de la séquence thématique la plus probable pour un texte permet d'attribuer un auteur à chaque phrase.

Abstract In this contribution, we show how we used probabilistic methods from unsupervised text mining to solve a supervised task: the DÉfi Fouille de Textes 2005. Our model consists of a mixture of multinomial distributions over the word counts, each component corresponding to a different theme. Each theme is associated with one author. We apply the EM algorithm to estimate the parameters of these thematic distributions. In the testing phase, we define one latent variable per sentence and link the themes within a document with a hidden Markov model with fixed parameters. Finding the likeliest state sequence finally enables us to attribute every sentence to its most likely author.

1 Introduction

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 (LSA, Deerwester *et al.*, 90)), des méthodes probabilistes ont trouvé leur place pour l'analyse exploratoire de données textuelles, les modèles plus populaires étant probablement *Probalistic latent semantic analysis* (PLSA, Hofmann, 01) et *Latent Dirichlet Allocation* (LDA, Blei *et al.*, 02). À l'instar de ces auteurs, nous plaçons ici dans le domaine des statistiques paramétriques et utilisons un modèle de mélange dont les variables latentes ont une interprétation *thématique*. Les paramètres de ces modèles ont une 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.

Pour contourner les inconvénients de ces modèles, essentiellement liés à leur complexité, nous considérons ici un modèle plus simple (Nigam *et al.*, 00; Clérot *et al.*, 04), dans lequel chaque document est supposé monothématique. 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 utilisée pour le "DÉfi Fouille de Textes" (DEFT).

Dans la deuxième partie de l'article, nous expliquons comment utiliser ce modèle pour DEFT, en identifiant des thèmes dans les discours de chaque locuteur. Nous introduisons une variable latente de thème par *phrase*. Le lien entre la variable indicatrice du thème d'une phrase et celle de la phrase suivante est réalisé par un modèle de Markov caché, dont les paramètres sont supposés connus. L'algorithme de Viterbi permet ensuite de proposer une séquence d'états la plus vraisemblable et, par suite, l'auteur probable de chaque phrase.

Finalement, nous étudierons les résultats obtenus par le modèle proposé et ses variantes avant de conclure sur les améliorations possibles et travaux à venir.

2 Modèle de mélange de multinomiales

2.1 Préliminaires et notations

Pour représenter les textes, nous adoptons le modèle du sac-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. On note n_D, n_D^*, n_T et n_W respectivement les nombres de documents dans les corpus d'apprentissage et de test, le nombre de thèmes (i.e. le nombre de composantes du modèle de mélange dans la section 2.2) et la taille du vocabulaire.

Pour $d \in \{1, \ldots, n_D\}$, $d^* \in \{1, \ldots, n_D^*\}$ et $w \in \{1, \ldots, n_W\}$, on note $C_d(w)$ et $C_{d^*}^*(w)$ les termes généraux des matrices de comptes d'entraînement et de test, c'est-à-dire les nombres d'occurrences du mot w dans les documents numéros d et d^*. On note également $l_d = \sum_{w=1}^{n_W} C_d(w)$ le nombre de mots dans le texte d et $l = \sum_{d=1}^{n_D} l_d$ le nombre total de mots dans le corpus d'entraînement, somme de tous les termes de la matrice de comptes. On définit similairement $l_{d^*}^*$ et l^* pour le corpus de test.

2.2 Modèle génératif

Contrairement au cadre de la catégorisation supervisée, on considère ici que l'on ne dispose d'aucune information sur les classes. Le modèle de génération du corpus que nous présentons nous permet, après estimation des paramètres, de proposer un classement des documents suivant les différentes composantes du mélange.

On suppose que les textes sont indépendants. Chaque document (numéroté $d \in \{1, \ldots, n_D\}$) 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. D'où le modèle génératif pour un document:

- Tirer un thème $t \sim \text{Mult}(1, (\alpha_1, \ldots, \alpha_{n_T}))$[1] où les $\alpha_{t'}$ sont des paramètres tels que $\sum_{t'=1}^{n_T} \alpha_{t'} = 1$.
- Tirer l_d mots $C_d = (C_{d1}, \ldots, C_{dn_W}) \sim \text{Mult}(l_d, (\beta_{1t}, \ldots, \beta_{n_W t}))$, 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:

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

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. On fait appel à l'algorithme EM (Expectation Maximization) dans lequel on s'intéresse à l'espérance, conditionnellement aux observations, de la log-vraisemblance *complète* $\mathcal{L}^c$, c'est-à-dire la log-vraisemblance des couples (vecteurs de comptes, thème) en supposant que le thème correspondant au texte d est t_d.

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

[1] On note $\text{Mult}(k, (\alpha_1, \ldots, \alpha_{n_T}))$ l'opération consistant à tirer k fois suivant une multinomiale de probabilités $(\alpha_1, \ldots, \alpha_{n_T})$.

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:

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

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. Donc, pour $t \in \{1, \ldots, n_T\}$ et $w \in \{1, \ldots, n_W\}$:

$$\alpha_t = \frac{1}{n_D} \sum_{d=1}^{n_D} p(T_d = t|C_d; \alpha', \beta') \quad (2)$$

$$\beta_{wt} = \frac{\sum_{d=1}^{n_D} C_d(w) p(T_d = t|C_d; \alpha', \beta')}{\sum_{w=1}^{n_W} \sum_{d=1}^{n_D} C_d(w) p(T_d = t|C_d; \alpha', \beta')} \quad (3)$$

Ces formules sont appliquées 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. Dans la suite, nous utilisons un lissage de Laplace, consistant à augmenter tous les comptes de 0.1, ce qui revient à mettre sur les paramètres β une distribution *a priori* suivant une loi Dirichlet de paramètre 1.1.

2.3 Méthode de construction 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.*, 05), dont nous résumons ici les conclusions qui nous semblent pertinentes pour DEFT. Le corpus qui a servi de base à ces expérimentations est un corpus raisonnablement simple, issu de Reuters 2000[2] et composé de 5000 textes équirépartis dans 5 catégories (arts, sports, emploi, catastrophes, santé). En plus de la log-vraisemblance, nous considérons deux mesures des performances:

- La *perplexité*, qui quantifie la capacité du modèle à prédire de nouvelles données.
- *L'information mutuelle* entre le classement produit par le modèle et les catégories Reuters pré-existantes. 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 l'obtention de 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 construire à partir de regroupements qui se recoupent largement, l'apprentissage se chargeant en général de les séparer. Dans cet esprit, l'algorithme est initialiser en fixant les probabilités *a posteriori* pour un document d'appartenir à un thème, équation (1), très proches de l'équiprobabilité entre tous les thèmes. Pour chaque essai, on tire donc ces valeurs selon une distribution Dirichlet de variance faible.

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

Ces expériences nous ont conduit à établir les constats suivants:

- La variabilité entre les deux initialisations est très forte pour toutes les mesures, log-vraisemblance, perplexité et information mutuelle.
- La log-vraisemblance est un indicateur raisonnable de la qualité finale du regroupement produit,; c'est le seul indicateur que l'on puisse obtenir dès la phase d'apprentissage.
- 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.

Sur la base de ces observations, l'idée de la méthode d'initialisation finalement retenue est la suivante: partant d'un vocabulaire extrêmement réduit (environ 1000 mots, soit 2% du vocabulaire total) avec l'initialisation "Dirichlet", une première estimation des paramètres du modèle est obtenue. Ce procédé est répété plusieurs fois et seul le meilleur ensemble de paramètres (au sens de la log-vraisemblance finale) est conservé. La taille du vocabulaire est ensuite progressivement augmentée, en réinitialisant à chaque étape le modèle sur les probabilités *a posteriori* issues de l'étape précédente. Cette procédure est itérée jusqu'à ce que le vocabulaire complet soit finalement pris en compte.

[2]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.

Les résultats présentés dans (Rigouste *et al.*, 05) montrent que l'algorithme d'initialisation itératif parvient au final 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 valeurs beaucoup plus satisfaisantes qu'avec l'initalisation "Dirichlet" simple.

3 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.

3.1 Prétraitements

Pour chacune des trois tâches, la même série de pré-traitements des corpus a été utilisée, consistant à segmenter chaque phrase en mot, à normaliser les chiffres, à mettre tous les mots en minuscule et à supprimer toutes les marques de ponctuation.

À l'issue de ces traitements, le vocabulaire utilisé dans le modèle statistique 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.

On suppose 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 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.

3.2 Description de l'algorithme

L'algorithme itératif d'estimation des paramètres décrit en section 2.3 est utilisé pour obtenir les coefficients β_{wt} correspondant aux thèmes récurrents des discours de J. Chirac. Le nombre de thèmes n_{T_C} est fixé *a priori*. On procède de même pour F. Mitterrand, avec un nombre de thèmes n_{T_M}, permettant d'obtenir au total $n_{T_C} + n_{T_M}$ 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, on utilise un modèle de Markov caché dont les probabilités de transition sont supposées connues. Pour chaque document du corpus de test, l'"état" (c'est-à-dire le thème) le plus probable de chaque phrase est déterminé par application de l'algorithme de Viterbi.

3.3 Prise en compte des contraintes

L'algorithme précédent ne permet pas de respecter différentes contraintes qui constituent pourtant des informations intéressantes:

1. Un texte commence toujours par un thème "J. Chirac".
2. Toute transition directe entre deux thèmes du même locuteur est interdite (à l'exception des insertions, chaque texte est donc supposé monothématique).
3. Chaque fragment contient toujours au moins deux phrases (pas de phrases de F. Mitterrand isolées et au moins deux phrases de J. Chirac en début de texte).
4. L'insertion d'un fragment "F. Mitterrand" sépare deux fragments "J.Chirac" qui appartiennent au même thème.
5. Il n'y a qu'une seule insertion "F. Mitterrand" par document.

Les deux dernières conditions ne sont pas explicites dans les règles de DEFT. Après examen rapide du corpus, il nous a cependant semblé que ces idées simples devraient permettre d'obtenir de meilleurs scores. Pour le vérifier, nous testons donc 3 modèles: le modèle 1 ne tient pas compte des deux dernières contraintes ; le modèle 2 respecte la contrainte 4 mais pas 5; le modèle 3 suit toutes les contraintes ci-dessus.

L'incorporation de ces contraintes dans le modèle s'effectue en dupliquant les états de la chaîne de Markov et en adaptant au besoin les probabilités de transition. Ainsi le modèle 1 correspond à la machine à états finis représentée en figure 1, à gauche, dans l'hypothèse où $n_{T_C} = n_{T_M} = 2$. Pour obtenir le modèle 2, il faut dupliquer tous les états "F. Mitterrand" pour chaque thème de "J. Chirac", comme indiqué à droite sur la même figure.

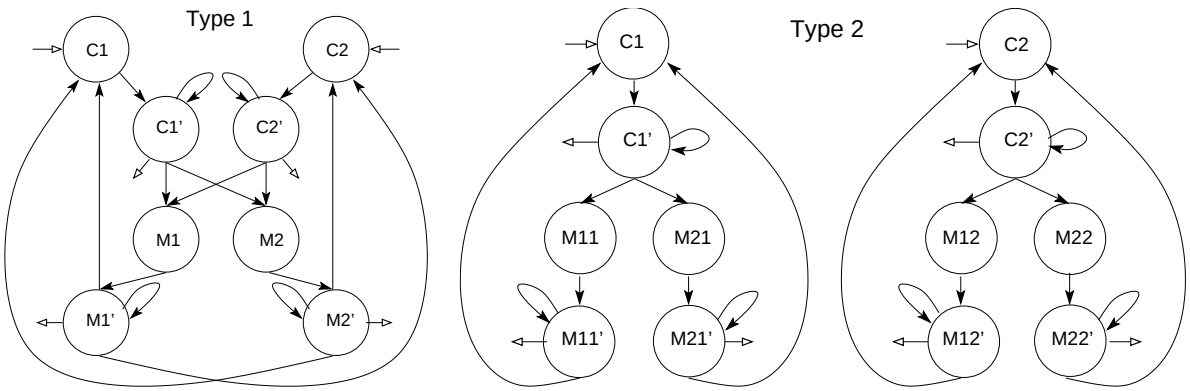

Figure 1: Modèles 1 et 2.

Enfin, dans le cas du modèle 3 (figure 2), on doit également dupliquer les thèmes “J. Chirac” pour avoir un état “pré-insertion” et un état “post-insertion”.

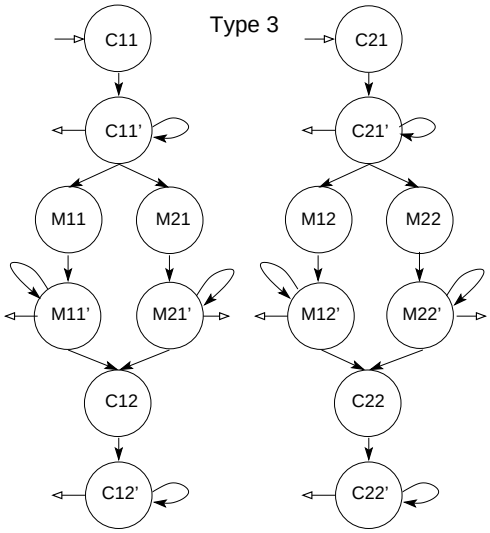

Figure 2: Modèle 3.

Les probabilités de transition et les probabilités de sortie, non figurées sur les graphes précédents, sont calculées comme suit. Pour les probabilités de transition, on multiplie les constantes de changement d’auteur (p_{C2M} et p_{M2C}, fixées à 0.3) par le paramètre α_t correspondant au thème dans lequel on entre. Les probabilités de sortie ont été en général fixées à 0, à l’exception notable du modèle 3 où les probabilités de sortie des états “J. Chirac” 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 dans le même état (boucle) sont calculées pour que la somme des probabilités de transition pour un état donné soit égale à 1.

3.4 Résultats

Nous étudions ici uniquement les résultats sur la tâche 1 de DEFT[3]. Pour la campagne officielle de test, nous avons soumis les modèles 1 et 2 avec $n_{T_C} = 10$ et $n_{T_M} = 4$, le modèle 2 obtenant les meilleures performances. La figure 3 montre qu’il est possible d’atteindre des performances légèrement meilleures en utilisant le modèle 3. Le meilleur résultat obtenu à ce jour, de 0.88, est obtenu avec le modèle 3 en fixant $n_{T_C} = 10$ et $n_{T_M} = 3$.

Dans tous les cas, il semble que les nombres optimaux de thèmes soient de 3 pour “F. Mitterrand” et de 8 ou 10 pour “J. Chirac”. Cette différence s’explique par la quantité de données d’apprentissage, bien plus importante pour un locuteur que pour l’autre, et qui permet par conséquent d’estimer de façon fiable un plus grand nombre de paramètres.

Enfin, pour évaluer l’apport de l’algorithme itératif d’initialisation utilisé pour apprendre les paramètres des thèmes, nous l’avons comparé avec les performances obtenues en utilisant une

[3]N’ayant pas cherché à tirer profit des informations spécifiques liées aux noms et aux dates, nos performances sur les tâches 2 et 3 sont quasiment les mêmes que sur la tâche 1.

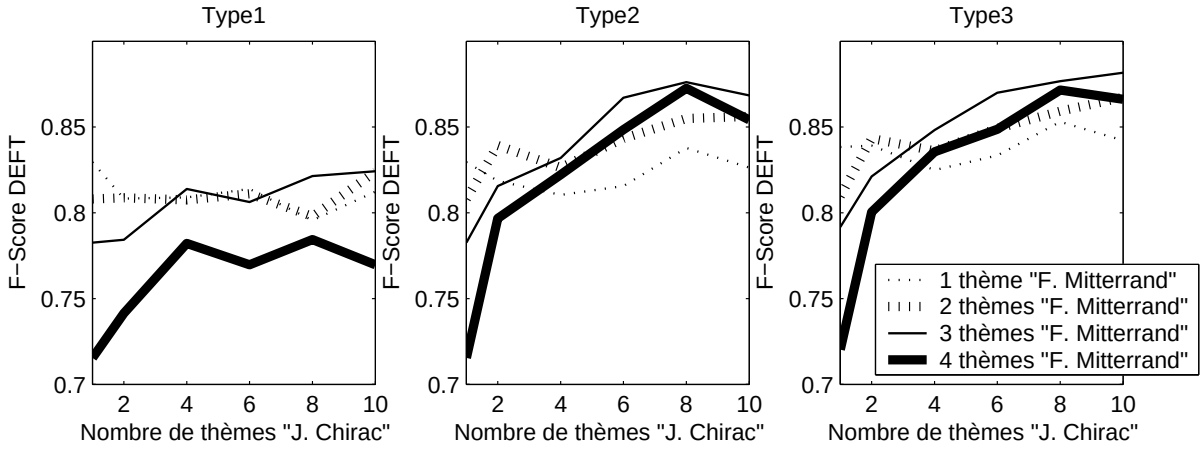

Figure 3: F-Score obtenu sur la tâche 1 de DEFT en fonction de n_{T_C}

procédure d'initialisation plus simple (initialisation "Dirichlet"), qui considère d'emblée tout le vocabulaire. Comme le montrent les résultats de la table 1, moyennés sur 50 tirages, les performances nettement meilleures obtenues par la méthode itérative sur la mesure "perplexité" se traduisent également un gain, d'ampleur plus modeste, sur la tâche d'évaluation extrinsèque de DEFT.

Méthode	Perplexité - corpus C	Perplexité - corpus M	DEFT F-Score
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

Table 1: Résultats comparés pour deux méthodes d'inférence des paramètres

4 Conclusion

Nous avons présenté dans cet article la méthode utilisée pour répondre au problème posé dans le cadre du DÉfi Fouille de Textes 2005. Il s'agit d'utiliser un modèle non supervisée d'analyse exploratoire pour une tâche de fouille de textes supervisée. 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 catégoriser les phrases du corpus de test, en déterminant l'enchaînement thématique le plus probable.

Le fait que notre modèle n'ait pas été spécifiquement conçu pour ce travail mais y obtienne malgré tout des résultats satisfaisants démontre son efficacité à segmenter un discours en thèmes, quand bien même les thèmes obtenus sont parfois difficile à analyser. Sur la base des résultats disponibles à ce jour, il semble que comparativement aux autres méthodes, notre modèle est plus efficace pour la tâche 1 que dans les deux autres. Pour obtenir de meilleures performances sur les tâches 2 et 3, il aurait fallu ajuster les poids des dates et des noms de personnes dans le calcul de la vraisemblance de chaque phrase.

Dans le cadre des travaux à venir, nous prévoyons de tester sur d'autres tâches la combinaison de ce modèle de mélange thématique et d'un modèle de Markov caché sur l'enchaînement des variables latentes associées aux phrases ou aux paragraphes. Une autre direction intéressante de

recherche consiste à comparer la méthode itérative heuristique d'inférence présentée dans cet article avec des algorithmes plus évolués de type échantillonneur de Gibbs.

Remerciements

Ce travail est financé par France Télécom, Division R&D, sous le contrat n°42541441.

Références

Blei D., Ng A., Jordan M. (2002), Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems (NIPS)*, Vol. 14, 601-608.

Clérot F., Collin O., Cappé O., Moulines E. (2004), Le Modèle "Monomaniaque" : un Modèle Statistique Simple pour l'Analyse Exploratoire d'un Corpus de Textes, *Colloque International sur la Fouille de Texte (CIFT'04)*.

Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R. (1990), Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, Numb. 6, 391-407.

Hofmann T. (2001), Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning Journal*, Vol. 42, Numb. 1, 177-196.

Nigam K., McCallum A., Thrun S., Mitchell T. (2000), Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, Vol. 39, Numb. 2/3, 103-134.

Rigouste L., Cappé O., Yvon F. (2005), Inference for Probabilistic Unsupervised Text Clustering, soumis à SSP 2005, IEEE Workshop on Statistical Signal Processing.