Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac

Marc El-Bèze, Juan-Manuel Torres-Moreno, Frédéric Béchet

LIA – Université d'Avignon et des Pays de Vaucluse
BP 1 228, F-84 911 Avignon Cedex 09
{ marc.elbeze, juan-manuel.torres, frederic.bechet }@univ-avignon.fr

Mots-clés : Segmentation, Catégorisation thématique, Adaptation, Cohésion

Keywords: Segmentation, Topic Classification, Adaptation, Cohesion

Résumé Nous présentons des modèles d'apprentissage probabilistes appliqués à la tâche de classification binaire telle que définie dans le cadre du défi DEFT05[1]. Au sein de discours de Jacques Chirac, a pu être insérée une séquence de phrases de François Mitterrand. Pour identifier la paternité de ces séquences, nous avons utilisé des chaînes de Markov, des modèles bayesiens, et des procédures d'adaptation de ces modèles. Une comparaison avec diverses approches montre la supériorité des méthodes que nous proposons. Les résultats que nous obtenons, en termes de précision, rappel et Fscore sur le sous corpus de test Mitterrand sont très encourageants.

Abstract We present probabilistic learning models applied to the task of binary classification as defined in the DEFT05[1] challenge (a sequence of François Mitterrand's sentences could have been inserted into a speech of Jacques Chirac). Markov chains, Bayes models and an adaptative process have been used. A comparison with a baseline and perceptron approaches shows the superiority of our methods. In terms of precision, recall and Fscore over the Mitterrand test sub-corpus our results are very promising.

[1] http://www.lri.fr/ia/fdt/DEFT05

1 Introduction

A priori, un travail de classification à 2 classes (ici Chirac et Mitterrand[2]) paraît simple. Or, de nombreuses raisons font que le problème est complexe. Au terme d'une étude portant sur 68 interventions télévisées composées de 305 124 mots, (Labbe, 1990) distingue 4 périodes dans les discours de Mitterrand. L'une d'elles dénommée « *le président et le premier ministre* » (octobre 1986-mars 1988) n'est probablement pas la plus facile à traiter sous l'angle de vue particulier proposé par le défi DEFT05. Dans d'autres conditions, c'eût été loin d'être évident. Ici, on peut s'attendre à des difficultés accrues pour différencier deux orateurs qui se sont exprimés dans maints débats sur les mêmes sujets. Facteur aggravant : on ne dispose que d'un petit corpus déséquilibré : 109 279 mots pleins pour l'un et 582 595 pour le second répartis dans 587 discours (dont la date n'est pas fournie). Notons qu'une classification supervisée binaire avec un perceptron optimal à recuit simulé (TORRES *et al.*, 2002) appliqué sur la catégorie grammaticale de mots (l'utilisation de tous les mots générant une matrice trop volumineuse) donne un taux d'extraction des segments Mitterand décevant Fscore ≈ 0.43 ; la méthode[3] K-means sur les mêmes données conduit à un Fscore ≈ 0.4. Avec des classifieurs à large marge (AdaBoost avec BoosTexter et SVM avec SVM-Torch), on plafonne à 0.5.

Dans cet article, nous décrivons quelques méthodes employées dans le cadre de ce défi. Nous présentons en section 2 une première approche reposant sur des modèles bayésiens, une chaîne de Markov, des adaptations statiques et dynamiques et un réseau sémantique de noms propres. En section 3, nous présentons une deuxième approche probabiliste ne faisant appel à aucun filtrage ou lemmatisation, combinée avec un automate légèrement différent. Des expériences et résultats sont présentés en section 4. Une méthode de fusion de plusieurs approches y est esquissée avant de conclure et d'envisager quelques perspectives.

2 Modélisation I

La chaîne de traitement que nous allons décrire dans les sous-sections suivantes est constituée de 4 composants dont un seulement est totalement dédié à la tâche de DEFT05. On pourrait facilement le modifier ou au pire s'en passer, s'il fallait changer de domaine d'application, Sur un Pentium portable cadencé à 1,7 GHz et doté d'une RAM de 384Mo, l'intégralité de la chaîne s'exécute en 20' qui se décomposent en 5' pour l'apprentissage, et 15' pour le test soit une minute par itération du couple adaptation – étiquetage.

2.1 Modèles Bayésiens

Guidée par une certaine intuition que nous avions des caractéristiques de la langue et du style de chacun des deux orateurs, une analyse des données d'apprentissage nous a poussés à retenir

---

[2] Pour des facilités d'écriture, nous prenons la liberté de désigner les deux derniers présidents de la République, par leur nom de famille, sans les faire précéder d'un titre, ou d'un prénom, et pour plus de concision, il nous arrivera de nous contenter de remplacer Mitterrand et Chirac par les étiquettes *M* et *C*.

[3] En comparaison, avec une méthode de type *base-line (random)* où avec une probabilité de 0.80 pour la classe *C* et 0.20 pour *M*, on assigne la classe d'une phrase du test, on obtient un Fscore ≈ 0.22 .

certaines de leurs caractéristiques plutôt que d'autres. En premier lieu, il était naturel de tabler sur une caractérisation s'appuyant sur les différences de vocabulaire. Des études anciennes comme celles de (COTTERET & MOREAU, 1969) sur le vocabulaire du Général de Gaulle, ou d'autres plus récentes (LABBE, 1990) partent du même présupposé. Pour plusieurs raisons, cette approche est incontournable mais comme on en rencontre tôt ou tard les limites, on est amené naturellement à ne pas s'en contenter. En effet, la couverture des thématiques abordées par les différents présidents est très large. Les trajets politiques de deux présidents consécutifs se recoupent forcément. En conséquence, on observe de nombreux points communs dans leurs interventions, recouvrements auxquels viennent s'ajouter les reproductions conscientes ou inconscientes (citations ou effets de mimétisme).

Pour diversifier les points d'appuis, nous en avons testé d'autres comme la longueur des phrases (LL), le pourcentage de conjonction de subordination (Pcos), d'adverbes (Padv) ou d'adjectifs (Padj). Cinq de ces variables (Pcos, Padv, Padj, LL , et Plm) ont été modélisées par des gaussiennes dont les paramètres ont été estimés sur le seul corpus d'apprentissage. En ce qui concerne, le vocabulaire lui-même, qu'il s'agisse de lemmes ou de mots, nous avons entraîné sur ce même corpus des modèles *n*-grammes et *n*-lemmes (P#M et P#L), avec $n < 3$.

$$P(t) = \lambda_0 \times p_0(t) + (1-\lambda_0)\sum_{i=1}^{n} \lambda_i \times p_i(t) \qquad avec \ \sum_{i=1}^{n} \lambda_i = 1 \quad (1)$$

Les valeurs des coefficients λ_i que nous avons attribuées de façon empirique à chacune de ces 9 variables figurent dans le tableau 1 ci-dessous.

	P1L	P1M	Padj	LL	P2L	P2M	Pcos	Plm	Padv
λ_i	0.39	0.15	0.15	0.14	0.05	0.04	0.05	0.02	0.01

Tableau 1 : Caractères employés pour la modélisation bayésienne et coefficients associés

Lorsqu'on utilise des chaînes de Markov en TALN, on est toujours confronté au problème de la couverture des modèles. Le taux de couverture décroît quand augmente l'ordre du modèle. Le problème est bien connu et des solutions de type lissage ou Back-off (MANNING & SCHÜTZE, 2000) sont une réponse classique au fait que le corpus d'apprentissage ne suffit pas à garantir une estimation fiable des probabilités. Le problème devenant critique lorsqu'il y a un déséquilibre flagrant entre les deux classes, il nous a semblé inutile, voire contre-productif de calculer des trigrammes sur le sous corpus *M*.

En nous inspirant des travaux menés en lexicologie sur les discours de Mitterrand, nous avons essayé de prendre en compte certains des traits qualifiés de dominants chez Mitterrand par (ILLOUZ *et al.,* 2000) : adverbe négatif, pronom personnel première personne singulier, point d'interrogation, ou des expressions comme *c'est, il y a, on peut, il faut* (dans les 4 cas, à l'indicatif présent). Après vérification de la validité statistique de ces traits sur le corpus DEFT05, nous les avons intégrés dans la modélisation mais dans un second temps, nous les avons retirés car même s'ils entraînaient une légère amélioration sur les données de développement, rien ne garantissait qu'il ne s'agissait pas, là, de tics de langage liés à une période potentiellement différente de celle du corpus de test. Par ailleurs, en cas de portage de l'application à un autre domaine ou une autre langue, nous ne voulions pas être dépendants d'études lourdes. En tous les cas, nous avons préféré faire confiance aux modèles de Markov pour capturer automatiquement une grande partie de ces tournures.

2.2 Prise en compte de contraintes générales

Un discours de Chirac peut avoir fait l'objet de l'insertion d'au plus une séquence de phrases. La séquence *M*, si elle existe, est d'une longueur supérieure ou égale à deux. Pour prendre en compte cette contrainte particulière, nous avions, initialement, pensé écrire des règles, même si une telle façon de faire s'accorde généralement peu avec les méthodes probabilistes. Dans le cas présent, que faut-il faire si une phrase détachée de la séquence *M* a été étiquetée *M*, avec une probabilité plus ou moins élevée (certainement au dessus de 0.5, sinon elle aurait reçu l'étiquette *C*) ? Renverser la décision, ou la maintenir ? Si l'on opte pour la seconde solution, il serait logique d'extraire également toutes les phrases qui la séparent de la séquence *M*, bien qu'elles aient été étiquetées *C*. Ne pas se contenter du *statut quo* comporte un piège : un gain aléatoire en rappel risque de se faire au prix d'une chute de précision.

Pour pouvoir trouver, parmi les chemins allant du début à la fin du discours, celui qui optimise la production globale du discours, nous avons exploité un automate probabiliste à cinq états (dont un initial *I* et 3 terminaux, C_1, C_2, et M_2). Comme on peut le voir sur la figure 1, vers les états dénommés C_1 et C_2 (resp. M_1 et M_2), n'aboutissent que des transitions étiquetées *C* (resp. *M*). À une transition étiquetée *C* (resp. *M*), est associée la probabilité d'émission combinant pour *C* (resp. *M*) les modèles probabilistes définis en section 2.1.

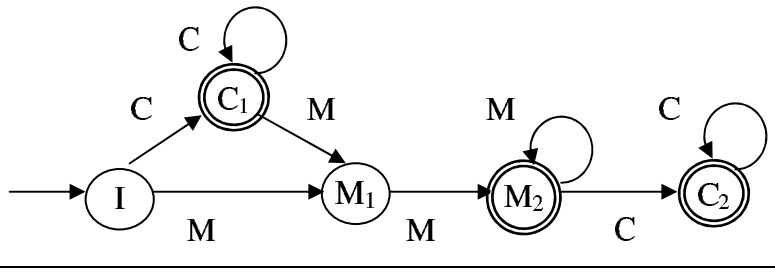

Figure 1 : Machine de Markov exprimant les contraintes générales des trois tâches

Avant de décrire les étapes ultérieures du processus de catégorisation segmentation, notons que c'est ce composant qui a permis de faire un saut conséquent (plus d'une dizaine de points) au niveau des performances et a ouvert ainsi la voie à la mise en place de procédures d'adaptation décrites en section suivante. Remarquons par ailleurs que la question aurait pu être gérée autrement, par exemple en utilisant, pour chaque discours, la partie triangulaire d'une matrice carrée *M*[*n*,*n*] (*n* étant le nombre de phrases contenues dans le discours en question). Dans chaque case *M*[*i*,*j*], on calcule la probabilité que la séquence soit étiquetée *M* entre *i* et *j*, et *C* du début jusque *i-1* et de *j+1* à *n*. Déterminer les bornes optimales de la séquence Mitterrand revient alors à rechercher un maximum sur toutes les valeurs *M*[*i*,*j*] telles que $i > j$. Si celle valeur optimale est inférieure à celle qu'on aurait obtenue en produisant toute la chaîne avec le modèle associé à Chirac, on se doit de supprimer la séquence *M*.

La complexité de cette seconde méthode est supérieure à celle de l'algorithme de Viterbi que nous avons employé. Il nous a paru néanmoins intéressant d'en faire état car elle offre la possibilité de combiner aisément des contraintes globales plus élaborées que celles que nous avons à prendre en compte. Elle peut aussi permettre de mixer des modèles issus de l'apprentissage et d'autres optimisant des variables dédiées à la modélisation de la cohésion interne des séquences qui se trouvent dans le discours traité, et n'ont fait l'objet d'aucun apprentissage préalable.

2.3 Adaptation statique et dynamique

Durant cette étape, ont été mises en œuvre des procédures d'adaptation statique et dynamique qui permettent de gagner entre 3 et 4 points de Fscore. La contrainte de ne pouvoir enrichir le corpus d'apprentissage, sous peine de disqualification, nous a poussé à tirer un parti intégral des données mises à notre disposition. Or, en dehors du corpus d'apprentissage, ne restait plus que les données de test. C'est sur elles, que l'adaptation a été pratiquée. Dériver un modèle à partir de l'intégralité des données de test correspond à ce que nous appelons ici adaptation statique. L'adaptation dynamique, quant à elle, repose sur un modèle découlant seulement du discours en train d'être testé. Bien entendu, il n'est pas interdit de conjuguer les 2 approches.

Dans un premier temps, nous avions envisagé de pratiquer un étiquetage des données de test, l'objectif étant à l'itération $i+1$ de n'adjoindre au corpus d'apprentissage[4] de X que les phrases s ayant reçu au pas i une probabilité $P_i(X|s)$ supérieure à un certain seuil $T_{X,i}$. Un apprentissage de type maximum de vraisemblance effectué sur les données ainsi collectées peut autant rapprocher qu'éloigner du point optimal. Pour pallier cette difficulté, nous avons opté pour un apprentissage EM, consistant à ne compter pour chaque couple {élément=e, X} observé dans les données d'adaptation que la fraction d'unité égale à la probabilité de l'orateur X sachant la phrase qui contient e. La prise de décision repose sur une formule analogue à celle de la formule (1). La variable en position 0 est la probabilité de l'étiquette sachant la phrase qui lui a été attribuée à l'itération i. Nous avons fait décroître le poids λ_0 qui lui est associé, de façon progressive, d'une itération à l'autre par pas de 0.1. Les 4 modèles employés sont, pour les 2 premiers, lemmes et mots issus de l'adaptation locale, pour les 2 derniers, lemmes et mots issus de l'adaptation globale. La pondération entre les différentes probabilités est restée la même durant toutes les itérations : { `0.9, 0.02, 0.003, 0.005` }.

2.4 Réseau de Noms Propres et Cohérence interne des discours

À partir de la tâche 2, l'ensemble des noms propres était dévoilé aux participants. Établir un lien entre différents éléments apparaissant dans des phrases même éloignées d'un discours donné, nous a paru être un bon moyen pour mettre en évidence une sorte de réseau sémantique permettant aux segments de s'auto regrouper autour d'un lieu, de personnes et de façon implicite d'une époque. Dans le cas de données bien séparables, plusieurs ensembles de noms ancrés dans une Histoire et une Géographie commune devraient former des composantes connexes (idéalement deux) sur lesquelles il suffirait ensuite de mettre l'étiquette M ou C. Bien que cela ne soit pas tout à fait la démarche que nous avons adoptée, ces remarques aident à en comprendre l'esprit.

1339 termes ont été regroupés dans 275 « concepts » qui pour épouser la richesse des discours traités dépassent largement un cadre restreint aux seules considérations géopolitiques (le Sport et la Culture sont souvent abordés lors de cérémonies de remises de médailles). Un terme peut se retrouver dans plusieurs classes, comme par exemple Miguel Angel Asturias, qui a été placé aussi bien dans la classe des guatémaltèques que dans celle des écrivains étrangers. Afin de mixer les relations entretenues entre les noms de pays, leurs habitants, les capitales, le

---

[4] X pouvant prendre ici les valeurs M ou C .

pouvoir exécutif, nous avons complété un réseau fourni par le Centre de Recherche de Xerox, en y rajoutant quelques relations issues des Bases de Connaissance que l'équipe TALN du LIA utilise pour faire fonctionner son système de Questions / Réponses (BELLOT *et al.*, 2003). Ci-dessous, figure un petit extrait de ce réseau non structuré :

ARGENTIN
Argentine Alfonsin Carlos_Menem Bioy_Casares Buenos_Aires Alfredo_Arias Jorge_Remes
MEXICAIN
Mexique Mexico Zedillo Zédillo Benito_Juarez Carlos_Fuentes Octavio_Paz FOX Fox Cancun Monterrey

Après 4 itérations, sur 57 301 phrases valides que comptait le corpus de développement (test : 27 120), 6 011 ont été regroupées en 906 groupes (test : 432 groupes de 2 907). Plus de 10% des segments se retrouvent donc dans des groupes, dont le cardinal moyen est d'environ 6,5. Le plus grand groupe contient 50 segments (test : 63). Seuls 16 groupes (test : 12) regroupent, de façon confuse, des étiquettes M et C. C'est le cas du discours 38, où la phrase 30 étiquetée M possède en commun Casablanca MAGHREB (en fait, il s'agissait du sommet de Casablanca) avec la phrase 173 étiquetée *C*, où Chirac fait état de ses récents voyages au Maroc. L'avantage d'un réseau probabiliste est que cette erreur n'est pas rédhibitoire. En effet, dans notre soumission, la phrase 30 a été correctement extraite et non la phrase 173. Cela ne fonctionne pas toujours aussi bien ! Dans le cas du discours 739, la séquence *C* et la séquence *M* ont en commun 2 « termes-concepts » (Espagne-Espagnol et Méditerranée-Méditerranéen). Il se trouve que la seconde confusion aurait pu être évitée si le *TGV Paris-Lyon-Méditerranée* dont parle Mitterrand n'avait pas fait l'objet d'une sur découpe au moment de la tokenisation. Mais cela n'aurait pas suffi, car avec l'aide de l'autre terme (*Espagne*) quatre phrases M 30, 35, 36 et 37 ont été regroupées par transitivité avec 12 phrases étiquetées *C* (1, 3, 6-17, 20-5, 27, 47). De fait, aucun segment du discours 739 n'a été extrait. Il est clair que nous sommes encore loin d'une représentation élaborée des relations entretenues entre des concepts et leur expression au travers de textes, mais le réseau que nous avons élaboré à peu de frais est un premier pas dans cette direction.

3 Modélisation II

Nous nous sommes demandé si la recherche des caractéristiques propres à un auteur pourrait être facilitée par le fait de ne pas filtrer ou éliminer quoi que ce soit des discours. Ainsi, nous avons fait l'hypothèse que l'utilisation répétée, voire exagérée de certains symboles de ponctuation ou l'emploi de termes ne servant qu'à assurer le bâti de la phrase, pouvait prétendre au statut d'indicateur fiable.

3.1 Modèle probabiliste

Pour ce deuxième modèle, nous sommes partis du principe que les techniques de *n*-grammes appliquées à des tâches de classification, pourraient se passer d'une phase préalable de lemmatisation ou de stemming, du rejet des mots-outils et de la ponctuation. Pour les systèmes *n*-grammes, (Jalam & Chauchat, 2002 ; Sahami, 1999) ont montré que les performances ne s'améliorent pas après stemming ou élimination des mots-outils. Dans cet esprit, nous avons laissé les textes dans leur état originel. Aucun prétraitement n'a été effectué, même si cette démarche a ses limites : par exemple, *Gasperi* et *Gaspéri* comptent pour des mots différents, qu'il y ait ou non erreur d'accent ; *premier* et *première* sont aussi comptabilisés séparément en absence de lemmatisation. Malgré cela, nous avons voulu donner

au modèle un maximum de chances de capturer des particularités de style (manies de ponctuation, sur ou sous emploi de subjonctifs, gérondifs, …) qui auraient été gommées après application de prétraitements comme la lemmatisation.

3.2 Adaptation naïve

Une idée naïve nous est venue à l'esprit. Les contraintes DEFT05 indiquent qu'il y a zéro ou au moins 2 phrases de Mitterrand insérées dans tout discours de Chirac. De ce fait, nous avons implanté la méthode simple d'absorption suivante : si une phrase *i* appartenant à la classe *M* est précédée et suivie des phrases *i*-1 et *i*+1 ∈ à la classe *C*, alors on transforme l'étiquette de la phrase *i* en *C*. Le cas opposé a été également pris en compte : des phrases de la classe *C* enveloppées par des phrases du type *M* seront donc absorbées comme *M*. Et cela pour tous les discours. Nous avons étiqueté la première et dernière phrase de chaque discours comme appartenant à la classe *C*. Même si elle est simple, cette méthode fait gagner entre 4 et 5 points de Fscore.

3.3 Adaptation par Viterbi

La méthode naïve fait progresser de quelques points, mais elle présente deux inconvénients majeurs : i) elle est trop dépendante des contraintes fixées par DEFT05, ii) elle a tendance à laisser, de façon indésirable, des îlots de la classe *M* au milieu des discours. Comment les éliminer ? Nous avons procédé de la même manière que dans la section 2.2, avec un automate légèrement différent. Nous avons donc construit l'automate probabiliste à cinq états représenté en figure 2, avec un état initial *I*, final *F* et 4 états intermédiaires C_1, M_1, M_2 et C_2. Comme dans le cas de l'automate de la figure 1, à une transition étiquetée *C* (resp. *M*), est associée la probabilité d'émission combinant pour *C* (resp. *M*) les modèles de *n*-grammes définis en section 3.1. Nous avons alors appliqué l'algorithme de Viterbi (MANNING & SCHÜTZE, 2000) pour trouver la séquence optimale. Nous obtenons, de cette façon, un Fscore de 0.818 sur l'ensemble de développement de la tâche 1.

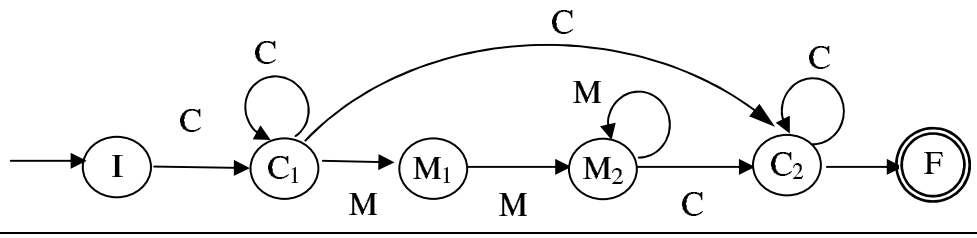

Figure 2 : Machine de Markov exprimant les contraintes générales des trois tâches

4 Expériences

Modélisation I : Tous les corpus (apprentissage et test) ont été traités par l'ensemble d'outils LIA_TAGG (www.lia.univ-avignon.fr), pour effectuer une tokenisation, un étiquetage morpho-syntaxique grâce au tagger dérivé du tagger ECSTA (SPRIET & EL-BÈZE, 1998). Dans la phase de développement, le corpus d'apprentissage a été découpé en 5 sous corpus de telle sorte que pour chacune des 5 partitions, un discours appartient dans son intégralité soit au test soit à l'apprentissage. À tour de rôle, chacun de ces sous corpus est considéré comme corpus

de test tandis que les 4 autres font office de corpus d'apprentissage. Cette répartition a été préférée à un tirage aléatoire des phrases tolérant le morcellement des discours. En effet, un tel tirage au sort présente deux inconvénients majeurs. Le premier provient du fait qu'un tirage aléatoire peut placer dans le corpus de test des segments très proches de segments voisins qui eux ont été placés dans le corpus d'apprentissage. Le second inconvénient (le plus gênant des deux), tient au fait qu'une telle découpe ne permet de respecter le schéma d'insertion défini dans les spécificités de DEFT05.

Modélisation II : Un autre protocole expérimental a été défini. Aucun filtrage, ni étiquetage syntaxique, ni lemmatisation. Nous avons créé des sous corpus d'apprentissage *A* et de développement *D* à partir du corpus d'apprentissage disponible (corpus.tache.learn), gardant la proportion de 80%-20% de discours respectivement. Nous avons éclaté le corpus d'apprentissage *A* en deux sous-ensembles, respectivement {*C*} et {*M*} contenant les phrases de Chirac ou celles de Mitterrand. Puis, nous avons construit les *n*-grammes (n = 1,2,3) de chaque sous-ensemble et nous avons calculé leur entropie moyenne. Nous avons obtenu $E(C) = 3.66$ et $E(M) = 3.84$. La classification des discours de l'ensemble de test *T* se fait comme suit : une phrase *i* d'un discours *j* est décomposée dans ses *n*-grammes, puis on calcule son entropie, celle de Chirac *EC* (sur les *n*-grammes de l'ensemble {*C*}) et celle de Mitterrand *EM* (sur les *n*-grammes de l'ensemble {*M*}). Nous avons défini un seuil $\delta = EC - EM$ et si la quantité $\delta < \varepsilon$ (avec ε suffisamment petit), la phrase sera attribuée à la classe Mitterrand, autrement à Chirac. Nous avons combiné par un lissage analogue à celui présenté en formule (1), avec λ_1=0.625, λ_2=0.166 et λ_3=0.208. Nous obtenons alors un Fscore de 0.83, 0.80.et 0.83 sur l'ensemble de développement des tâches 1, 2 et 3 respectivement.

4.1 Résultats

Pour alléger les graphiques, nous nous sommes limités au tracé de 4 courbes, par figure. Tant pour le Fscore que pour le rappel ou la précision, n'ont été retenus que les résultats obtenus sur les données de la tâche 2 (test et développement) dans 2 conditions expérimentales : avec ou sans les groupes de Noms Propres définis en section 2.4 (T ou D, avec ou sans Noms).

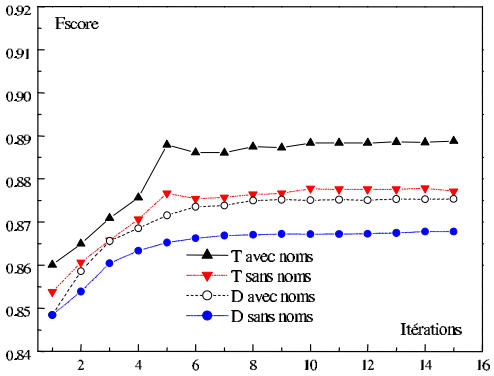

Figure 3 : Courbes de Fscore / tâche 2 / Modèle I / corpus de test (T) et développement (D)

Comme on peut le remarquer sur la figure 3, le Fscore s'améliore de façon notable au cours des 5 premières itérations. Au-delà, il n'y a pas à proprement parler de détérioration mais une stagnation qui peut être vue comme la captation par un maximum local. L'apport des réseaux bâtis autour des noms propres est indéniable, notamment c'est à eux que l'on doit le léger pic observé à la 5e itération.

4.2 Analyse des résultats

Ainsi qu'on peut le remarquer dans le tableau 2, récapitulant les résultats officiels de notre équipe, le dévoilement des dates (tâche T3) permet d'améliorer très légèrement les résultats du modèle **II**, mais entraîne une dégradation sur le modèle **I**. Il est intéressant de voir comment se comportent les courbes de précision et de rappel, au fil des itérations. La figure 4 le montre sur la tâche 2 : sur *T* ainsi que sur *D*, c'est le gain en précision qui explique l'amélioration due aux Noms Propres. Ce gain allant de pair avec un rappel quasi identique (légèrement inférieur pour le test), il apparaît que le composant Noms Propres fonctionne comme un filtre prévenant quelques mauvaises extractions (mais pas toutes).

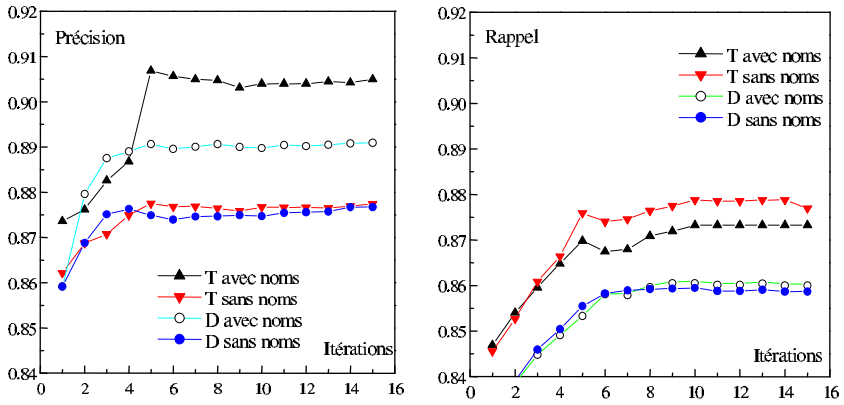

Figure 4 : Courbes de Rappel et Précision / tâche 2 / Modèle **I** / corpus *T* et *D*

4.3 Fusion de méthodes

Le dernier test que nous avons effectué (modèle **III**) repose sur une idée simple : accorder une confiance forte aux segments étiquetés de la même façon dans les différentes soumissions effectuées par les 2 équipes du LIA (junior et senior). Si cette méthode ne permet de gagner qu'entre 0 et 3 millièmes de point, il est clair qu'elle mériterait d'être appliquée sur des approches d'inspiration vraiment différente.

	Modèle I			Modèle II			Modèle III		
	Prec	Rappel	Fscore	Prec	Rappel	Fscore	Prec	Rappel	Fscore
T_1	0.881	0.854	0.867	0.826	0.764	0.794	0.883	0.858	0.870
T_2	0.909	0.861	**0.884**	0.827	0.775	0.8	0.911	0.858	**0.884**
T_3	0.887	0.872	0.879	0.829	0.776	0.801	0.89	0.871	0.880

Tableau 2 : Résultats officiels sur les 3 tâches {T_1, T_2, T_3} pour les trois soumissions

5 Conclusion et Perspectives

Les résultats que nous obtenons, en terme de Fscore (0.884) sont très encourageants**.** Ne pas lemmatiser et ne rien filtrer dégrade un peu les performances (Fscore≈0.84 avec le modèle **I**) mais permet de se passer d'un processus additionnel de prétraitement qui pour certaines langues peut être relativement lourd. Le recours à un réseau de Noms Propres est utile et nous encourage par la suite à employer une ressource lexicale comme EuroWordNet pour tirer parti

de réseaux sur les noms communs. Des frontières thématiques ne coïncident pas forcément avec des débuts de phrase. Les thèmes peuvent s'entremêler et composer un tissu discursif où les fils sont enchevêtrés de façon subtile. Beaucoup reste à faire pour pouvoir différencier plusieurs thèmes comme envisagé dans le cadre du Projet Carmel, plusieurs orateurs, ne serait-ce que trois. Et, si le lecteur veut se faire une petite idée de la difficulté de la tâche, nous l'invitons à deviner ce qui dans le présent article est dû à chacun de ses trois auteurs.

Remerciements

Nous remercions Eric Gaussier de Xerox d'avoir mis à notre disposition un lexique de Noms Propres. Nous sommes également reconnaissants envers Jérome Azé et Mathieu Roche du LRI qui n'ont pas ménagé leurs efforts pour organiser la campagne de DEFT05.

Références

BELLOT P., CRESTAN E., EL-BÈZE M., GILLARD L., DE LOUPY C. (2003), *Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11, Question-Answering Track*, actes de TREC'02, Gaithersburg, USA, NIST Special publication 500-251.

COTTERET J.-M., MOREAU R. (1969) *Le vocabulaire du Général de Gaulle,* Presses de la fondation nationale des sciences politiques, Armand Colin.

DAMASHEK M. (1995), *Gauging Similarity with N-Grams: Language-Independent Categorization of Text.* Science, 267 pp 843–848.

ILLOUZ G., HABERT B., FLEURY S., FOLCH H., HEIDEN S., LAFON P., PRÉVOST S. (2000), *Profilage de textes : cadre de travail et expérience*, Actes de JADT 2000, 5e Journées Internationales d'Analyse Statistiques des Données Textuelles, 9-11 Mars 2000, Lausanne.

JALAM R., CHAUCHAT J.-H. (2002), *Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques,* actes de JADT, 6e Journées Internationales d'Analyse Statistiques des Données Textuelles, pp 13-15, St-Malo.

LABBE D. (1990) *Le vocabulaire de François Mitterrand,* Paris, Presses de la Fondation Nationale des Sciences Politiques, mars 1990.

MANNING C. D., SCHÜTZE H. (2000) *Foundations of Statistical Natural Language Processing,* The MIT Press.

SAHAMI M. (1999), *Using Machine Learning to Improve Information Access.* PhD thesis, Computer Science Department, Stanford University.

SPRIET T., EL-BEZE M., (1998) *Introduction of Rules into a Stochastic Approach for Language Modelling*, Computational Models of Speech Pattern Processing, NATO ASI Series F, vol. 169, ed. Keith Ponting, pp. 350-355.

TORRES J.M., AGUILAR J.C., GORDON M.B. (2002), *Finding the number minimum of errors in N-dimensional parity problem with a linear perceptron.* Neural Processing Letters, p 201-210.