Un duel probabiliste pour départager deux Présidents

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

*Laboratoire d'Informatique d'Avignon - UAPV,
BP 1228, 84911 Avignon cedex 09, France
{ marc.elbeze, juan-manuel.torres, frederic.bechet }@univ-avignon.fr
http://www.lia.univ-avignon.fr
**École Polytechnique de Montréal - Département de Génie Informatique,
CP 6079 Succ. Centre-ville H3C 3A7 Montréal (Québec), Canada

Résumé. Nous présentons une palette de modèles probabilistes que nous avons employés dans le cadre du défi DEFT'05. La tâche proposée conjuguait deux problématiques distinctes du Traitement Automatique du Langage : l'identification de l'auteur (au sein de discours de Jacques Chirac, a pu être insérée une séquence de phrases de François Mitterrand) et la détection de ruptures thématiques (les thèmes abordés par les deux auteurs sont censés être différents). Pour identifier la paternité de ces séquences, nous avons utilisé des chaînes de Markov, des modèles bayésiens, et des procédures d'adaptation de ces modèles. Pour ce qui est des ruptures thématiques, nous avons appliqué une méthode probabiliste modélisant la cohérence interne des discours. Son ajout améliore les performances. Une comparaison avec diverses approches montre la supériorité d'une stratégie combinant apprentissage, cohérence et adaptation. Les résultats que nous obtenons, en termes de précision (0,890), rappel (0,955) et Fscore (0,925) sur le sous-corpus de test sont très encourageants.

1 Introduction

Dans le cadre des conférences TALN[1] et RECITAL[2] tenues en juin 2005 à Dourdan (France), un atelier a été organisé autour du défi de fouille textuelle proposé par (Azé et Roche, 2005). Ce défi portait le nom de DEFT'05 (DÉfi Fouille de Textes). Il a été motivé par le besoin de mettre en place des techniques de fouille de textes permettant soit d'identifier des phrases non pertinentes dans des textes, soit d'identifier des phrases particulièrement singulières dans des textes apparemment sans réel intérêt. Concrètement, il s'agissait de supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Cette tâche est proche de la piste *Novelty* du challenge TREC (Soboro, 2004) qui dans sa première partie consiste à identifier les phrases pertinentes puis, parmi celles-ci, les phrases nouvelles d'un corpus d'articles journalistiques. Pour mieux comprendre à quoi correspondait dans DEFT'05 la suppression des phrases non pertinentes d'un corpus de discours politiques (Alphonse et al.,

[1]Traitement Automatique des Langues Naturelles.
[2]Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues.

2005), une brève description du but général[3] s'impose. Un corpus de textes, allocutions officielles issues de la Présidence (1995-2005) de Jacques Chirac a été fourni. Dans ce corpus, des passages issus d'un corpus d'allocutions (1981-1995) du Président François Mitterrand ont été insérés. Les passages d'allocutions de F. Mitterrand insérés sont composés d'au moins deux phrases successives et ils sont censés traiter une thématique différente[4]. Un corpus, formé de discours de J. Chirac entrecoupés d'extraits de ceux de F. Mitterrand, est ainsi constitué.

Certaines informations sont supprimées de ce corpus afin de constituer les trois corpus ci-dessous :

- Corpus C1 : aucune présence d'années ni de noms de personnes : ils ont été remplacés par les balises <DATE> et <NOM> ;
- Corpus C2 : pas d'années : elles ont été remplacées par la balise <DATE> ;
- Corpus C3 : présence des années et des noms de personnes.

Le but du défi consistait à déterminer les phrases issues du corpus de F. Mitterrand introduites dans le corpus composé d'allocutions de J. Chirac. Ce but est commun aux trois tâches T1, T2 et T3 relatives aux trois corpus C1, C2 et C3 ont ainsi été définies. Intuitivement, la tâche T1 est la plus difficile des trois car le corpus afférent C1 contient moins d'informations que les deux autres. Les résultats (calculés uniquement sur les phrases de F. Mitterrand extraites) peuvent être évalués sur un corpus de test (T) avec des caractéristiques semblables à celui de développement (D) (cf. tableau 1), en calculant le *Fscore* :

$$Fscore(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel} \quad (1)$$

Dans le cadre de DEFT'05, le calcul du *Fscore* retenu par les organisateurs a été effectué uniquement sur les phrases de Mitterrand, et il a été modifié (Alphonse et al., 2005) comme suit (cette réécriture suppose évidemment que β soit égal à 1 de façon à ne privilégier ni précision ni rappel) :

$$Fscore(\beta = 1) = \frac{2 \times Nb\_phrases\_correctes\_extraites}{Nb\_total\_extraites + Nb\_total\_pertinentes} \quad (2)$$

- *Nb_phrases_correctes_extraites* : nombre de phrases qui appartiennent réellement au corpus de Mitterrand dans le fichier résultat fourni par le système ;
- *Nb_total_extraites* : nombre de phrases données dans le fichier résultat (que le système considère comme étant Mitterrand) ;
- *Nb_total_pertinentes* : nombre total de phrases de Mitterrand dans le corpus de test.

Discours	Phrases (D)	Mots (D)	Phrases (T)	Mots (T)
Chirac	52 936	1 148 208	30 148	638 547
Mitterrand	8 027	218 124	5 027	134 111

TAB. 1 – *Statistiques sur les corpus de développement (D) et de test (T).*

[3] Le lecteur intéressé trouvera une description plus détaillée ainsi que les données et résultats dans le site officiel du défi : http ://www.lri.fr/ia/fdt/DEFT05

[4] Par exemple, dans les allocutions de Jacques Chirac évoquant la politique internationale, les phrases de François Mitterrand introduites sont issues de discours traitant de politique nationale.

On pourrait être tenté de traiter chacune des trois tâches en appliquant les méthodes employées habituellement en classification. *A priori*, un problème de classification à deux classes (ici Chirac et Mitterrand[5]) paraît simple. Or, de nombreuses raisons font que la question est complexe. Au terme d'une étude portant sur 68 interventions télévisées composées de 305 124 mots, (Labbé, 1990) distingue quatre 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 particulier proposé par le défi DEFT'05. 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. Pour la tâche T1, 109 279 mots pleins pour un président et 582 595 pour le second répartis dans 587 discours (dont la date n'est pas fournie).

Pour donner une idée de la difficulté de ce défi, notons qu'une classification supervisée binaire avec un perceptron optimal à recuit simulé (Torres-Moreno 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 Mitterrand décevant avec un *Fscore* $\approx 0{,}43$; la méthode classique *K-means* sur les mêmes données, conduit à un *Fscore* $\approx 0{,}4$. En comparaison, avec des classifieurs à large marge réputés performants tels que *AdaBoost* (Freund et Schapire, 1997) avec *BoosTexter* (Schapire et Singer, 2000) et *Support Vector Machines* avec SVM-Torch (Collobert et Bengio, 2001), on plafonne à un *Fscore* $\approx 0{,}5$. Enfin, avec une méthode de type *base-line* vraiment simpliste, où on classerait tout segment comme appartenant à la catégorie *M*, on obtiendrait un *Fscore* $\approx 0{,}23$ sur l'ensemble de développement et de 0,25 pour le test. Comme ces résultats se sont avérés décevants, nous avons décidé d'explorer des voies totalement différentes. Nous présentons en section 2 une approche reposant sur des modèles bayésiens, une chaîne de Markov, des adaptations statiques et dynamiques et un réseau sémantique de noms propres. Des approches probabilistes avec ou sans filtrage et lemmatisation sont utilisées. En section 3, nous développons une approche de la cohérence interne des discours qui permet encore d'augmenter le *Fscore*. La section 4 est consacrée aux expériences et à leurs résultats. Nous comparons et tentons de fusionner ces différentes approches avant de conclure et d'envisager quelques perspectives. L'annexe présente une analyse détaillée d'un discours de la classe *C*, où l'utilisation de sa cohérence interne permet de mieux le classer.

2 Modélisation

La chaîne de traitement que nous allons décrire dans les sous-sections suivantes est constituée de quatre composants : un ensemble de modèles bayésiens (cf. 2.1), un automate de Markov (cf. 2.2), un modèle d'adaptation (cf. 2.3) et un réseau sémantique (cf. 2.4). Le seul composant totalement dédié à la tâche est l'automate. Le réseau sémantique dépend du domaine. Sur un Pentium portable cadencé à 1,7 GHz et doté d'une RAM de 384 Mo, l'intégralité de la chaîne d'adaptation 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. Le calcul de la cohérence interne demande un temps supplémentaire de 7' qui reste très raisonnable.

[5]Pour des facilités d'écriture, nous prenons dorénavant 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*.

2.1 Modèles bayésiens

Guidée par une certaine intuition que nous aurions pu avoir des caractéristiques de la langue et du style de chacun des deux orateurs, une analyse des données d'apprentissage aurait pu nous pousser à retenir certaines de leurs caractéristiques plutôt que d'autres. En premier lieu, il aurait été naturel de tabler sur une caractérisation s'appuyant sur les différences de vocabulaire. Des études anciennes comme celles de (Cotteret et Moreau, 1969) sur le vocabulaire du Général de Gaulle, ou d'autres plus récentes (Labbé, 1990) partent du même présupposé. Pour plusieurs raisons, cette approche semble prometteuse 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. Il est inévitable que les trajets politiques de deux présidents consécutifs se soient à maintes reprises recoupés. En conséquence, on observe de nombreux points communs dans leurs interventions. On suppose que ces recouvrements viennent s'ajouter les reproductions conscientes ou inconscientes (citations ou effets de mimétisme).

2.1.1 Modélisation avec lemmatisation

Au travers d'une modélisation classique (Manning et Schütze, 2000), nous avons testé quelques points d'appuis comme la longueur des phrases (LL), le pourcentage de conjonctions de subordination (Pcos), d'adverbes (Padv) ou d'adjectifs (Padj) et la longueur moyenne des mots pleins (Plm). Cinq de ces variables (Pcos, Padv, Padj, LL, et Plm) ont été modélisées par des gaussiennes p_i 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 $n < 3$. La probabilité de l'étiquette t (Chirac ou Mitterrand) résulte de la combinaison suivante :

$$P(t) = \sum_{i=1}^{r} \lambda_i \times p_i(t) \tag{3}$$

avec $\sum_{i=1}^{r} \lambda_i = 1$. Les valeurs des coefficients λ_i que nous avons attribuées de façon empirique à chacune de ces 9 variables i figurent dans le tableau 2. L'estimation de ces valeurs a bien entendu, été réalisée sur le corpus d'apprentissage. Comme le montre le tableau 2, le poids accordé aux lemmes est deux fois plus important que celui accordé aux mots.

i	P1L	P1M	Padj	LL	P2L	P2M	Pcos	Plm	Padv
λ_i	0,30	0,15	0,15	0,15	0,30	0,15	0,05	0,02	0,01

TAB. 2 – *Caractères employés pour la modélisation bayésienne et coefficients associés.*

Lorsqu'on utilise des chaînes de Markov en traitement automatique de la langue naturelle (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 et Schütze, 2000) ; (Katz, 1987) 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 tri-grammes.

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 à la première personne du singulier, point d'interrogation, ou des expressions comme « c'est », « il y a », « on peut », « il faut » (dans les quatre cas, à l'indicatif présent). Ceux-ci ont été traités de la même façon que les autres caractères de la modélisation bayésienne. Après vérification de la validité statistique de ces traits sur le corpus DEFT'05, 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.

2.1.2 Modélisation sans lemmatisation

Parallèlement et à l'inverse de nos préoccupations de la sous-section précédente, nous avons souhaité faire fonctionner nos modèles sur le texte à l'état brut, sans enrichissement ou annotation. Pour aller dans ce sens, nous nous sommes demandé à quel point la recherche automatisée des caractéristiques propres à un auteur pourrait être facilitée ou perturbée par le fait de ne pas filtrer ni éliminer quoi que ce soit des discours. Ainsi, nous avons fait l'hypothèse que l'utilisation répétée, voire exagérée de certains termes ne servant qu'à assurer le bâti de la phrase, pouvait prétendre au statut d'indicateur fiable. Pour ce deuxième modèle[6], 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. Les systèmes *n*-grammes, (Jalam et Chauchat, 2002) ; (Sahami, 1999) ont montré que leurs 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 ponctuer le texte par l'emploi de telle ou telle personne, de subjonctifs, gérondifs,...) qui sont gommées après application de certains prétraitements comme la lemmatisation. Une classification naïve et un calcul d'entropie ont déjà été rapportés lors de l'atelier DEFT'05 avec un automate légèrement différent (El-Bèze et al., 2005). Seule variante, l'ajout d'une contrainte : tout mot de longueur ≤ 5 n'est pas pris en compte afin d'alléger les calculs. Ceci correspond à un « filtrage » relativement indépendant de la langue.

[6] Qui sera appelé par la suite Modèle II et par opposition celui avec lemmatisation sera appelé Modèle I.

2.2 Automate de Markov

Comme cela était dit en introduction, 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*. Mais, dans ce cas, 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 trois 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 (respectivement M_1 et M_2) n'aboutissent que des transitions étiquetées *C* (respectivement *M*). À une transition étiquetée *C* (respectivement *M*), est associée la probabilité d'émission combinant pour *C* (respectivement *M*) les modèles probabilistes définis en section 2.1.

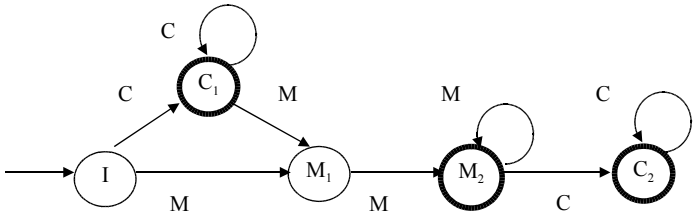

FIG. 1 – *Machine de Markov exprimant les contraintes générales des trois tâches.*

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 de 25% en absolu) au niveau des performances et a ouvert ainsi la voie à la mise en place de procédures d'adaptation décrites en section suivante. S'il s'avère qu'étiqueter un bloc de plusieurs segments est plus fiable qu'étiqueter individuellement chaque phrase, il est naturel que cela ait un impact positif sur les performances.

Remarquons par ailleurs que la question aurait pu être gérée autrement, par exemple en utilisant, pour chaque discours, la partie triangulaire supérieure d'une matrice carrée $\Psi[d, d]$ (d étant le nombre de phrases contenues dans le discours en question, voir les figures 2 et 3). Dans chaque case $\Psi[i, j]$, on calcule la probabilité que la séquence soit étiquetée *M* entre i et j, et *C* du début jusqu'au $i - 1$ et de $j + 1$ à d. Déterminer les bornes optimales de la séquence Mitterrand revient alors à rechercher un maximum sur toutes les valeurs $\Psi[i, j]$ telles que $i > j$. Si cette 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*. Sauf si on factorise les calculs pour remplir les différentes cases, la complexité de cette seconde méthode

est supérieure à celle de l'algorithme de Viterbi (Manning et Schütze, 2000). Il nous a paru néanmoins intéressant d'en faire état dès à présent, car elle offre la possibilité de combiner aisément des contraintes globales plus élaborées que celles que nous prenons en compte dans l'adaptation. 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, comme nous le montrerons en section 3.

2.3 Adaptation statique et dynamique

La contrainte de ne pouvoir enrichir le corpus d'apprentissage, sous peine de disqualification[7], nous a poussé à tirer un parti intégral des données mises à notre disposition. Or, en dehors du corpus d'apprentissage, il ne restait plus qu'une issue : intégrer dans l'apprentissage (bien entendu, sans les étiquettes de référence) une partie des données de test. C'est sur ces données que l'adaptation a été pratiquée. Dériver un modèle à partir de l'intégralité des discours 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. Évidemment, il n'est pas interdit de conjuguer les deux 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[8] de X que les phrases s ayant reçu au pas i une probabilité $P_i(X|s)$ supérieure à un certain seuil T_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 d'*Expectation-Maximisation* (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 (3). 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 quatre modèles employés sont, pour les deux premiers, lemmes et mots issus de l'adaptation locale (dynamique), pour les deux derniers, lemmes et mots issus de l'adaptation globale (statique). La pondération entre les différentes probabilités est restée la même durant toutes les itérations : Dynamique { lemmes = 0,4 ; mots = 0,1 } / Statique { lemmes = 0,4 ; mots = 0,1 }. Les procédures d'adaptation statique et dynamique mises en œuvre durant cette étape ont permis de gagner entre 3 et 4 points de *Fscore*.

2.4 Réseau de Noms Propres

À partir de la tâche T2, l'ensemble des noms propres était dévoilé aux participants. Établir un lien entre différents éléments apparaissant dans des phrases même éloignées d'un discours

[7] « Les équipes utilisant dans leur méthode des corpus de J. Chirac et de F. Mitterrand autres que ceux fournis par les organisateurs seront disqualifiées. Par exemple, la méthode consistant à acquérir un corpus de F. Mitterrand et/ou de J. Chirac par des ressources extérieures pour identifier les phrases de F. Mitterrand présentes dans le corpus fournis par les organisateurs sera considérée comme non valide. » Source : http ://www.lri.fr/ia/fdt/DEFT05/resultats.html

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

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.

2 171 termes ont été regroupés dans 314 « concepts »[9] 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 pouvoir exécutif, nous avons complété un réseau fourni par le Centre de Recherche de Xerox[10], 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). En table 3, figure un petit extrait de ce réseau non structuré. On enrichit les segments en leur

CONCEPT	TERMES
Argentin	Argentine Alfonsin Carlos_Menem Bioy_Casares Buenos_Aires Alfredo_Arias Jorge_Remes
Guatemalteque	Guatemala_Ciudad Guatemala Guatémaltèque Guatémaltèques Guatémaltais Guatémaltaise Guatémaltaises Ciudad_Vieja Permedo **Miguel_Angel_Asturias** Alvaro Arzu Irigoyen Rigoberta_Menchu Rigoberta
Mexicain	Mexique Mexico Zedillo Zédillo Benito_Juarez Carlos_Fuentes Octavio_Paz FOX Fox Cancun Monterrey Mexicain Mexicaine Mexicaines Mexicains
Ecrivains_e	Gao Saramago Virgilio_Ferreira Fernando_Pessoa Ionesco Cioran Fukuzawa_Yukichi Nadia_Tueni Amin_Maalouf Tahar_Ben_Jelloun Senghor Rachid_Boujedra Bioy_Casares Boubou_Hama Hector_Bianciotti **Miguel_Angel_Asturias** Dostoïevsky

TAB. 3 – *Extrait du réseau de noms propres.*

ajoutant les concepts auxquels appartiennent les termes qui les composent. Deux segments qui ont en commun plus d'un certain nombre d'éléments (termes ou concepts) sont considérés 2 à 2 comme liés thématiquement et mis dans une même classe de segments. Les classes sont élargies par une itération de ce processus. La probabilité de chaque segment est combinée avec la probabilité de la classe à laquelle il appartient. Après quatre itérations, sur 57 301 phrases valides que comptait le corpus de développement (test : 27 120), 6 285 ont été regroupées en 942 groupes (test : 456 groupes de 3 127). Un peu plus de 11% des segments se retrouvent donc dans des groupes, dont le cardinal moyen est d'environ 7 phrases. Le plus grand groupe

[9]Les termes ont été regroupés de façon manuelle pour former les concepts du réseau.
[10]http ://www.xrce.xerox.com

contient 50 segments (test : 66). 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 deux « 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, 20-25, 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. Néanmoins, le réseau que nous avons élaboré à peu de frais est un premier pas dans cette direction.

3 Cohésion thématique des discours

En section 2.2, nous avions avancé l'hypothèse qu'étiqueter un bloc est plus fiable qu'étiqueter chaque phrase de façon indépendante l'une de l'autre. Cela se discute en fait si on se borne à rechercher la suite de segments qui optimise la cohésion thématique[11] de chacun des deux blocs, il est indispensable de conjuguer cette approche thématique avec un étiquetage d'auteur. Cette étape est motivée par une idée simple découlant des présupposés de DEFT'05 : « *Les passages de F. Mitterrand introduits traitent d'une thématique différente. Par exemple, dans les allocutions de J. Chirac évoquant la politique internationale, les phrases de F. Mitterrand introduites sont issues de discours traitant de politique nationale. Ainsi, la rupture thématique peut être une des manières de détecter les phrases issues du corpus de F. Mitterrand.* »[12]

Dans cette optique, on peut vouloir trouver un découpage de chaque discours soit en un bloc (*C*...*C*), soit en deux blocs (*C* ... *C*-*M* ... *M*) ou (*M*...*M*-*C*...*C*) soit en trois blocs (*C*...*C*-*M*...*M*-*C*...*C*) tels que le bloc des segments étiquetés *M* ou les blocs (1 ou 2) étiquetés *C* présentent tous les deux une cohérence thématique interne optimale. Pour cela, nous proposons de formaliser le problème comme suit : la probabilité de production d'une phrase est évaluée au moyen d'un modèle appris sur toutes les phrases du bloc auquel elle appartient sauf elle. En maximisant le produit des probabilités d'émission de toutes les phrases du discours, on a toutes les chances de bien identifier des ruptures thématiques. Mais rien ne garantit qu'elles correspondent à des changements d'orateurs. En effet, supposons qu'il n'y ait pas dans un discours donné, d'insertion de phrases de Mitterrand, et que dans les discours de Chirac se trouve une longue digression de 20 phrases. Notre approche risque de reconnaître à tort ces 20 phrases comme attribuables à la classe *M*. Pour éviter ce travers, nous proposons une optimisation mettant en œuvre conjointement les modèles de cohérence interne et ceux issus de l'apprentissage.

[11] La cohésion thématique étant un des éléments permettant d'apprécier la cohérence interne d'un discours, nous emploierons de préférence l'expression « cohérence interne » dans la suite de l'article.

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

En annexe, nous donnons en exemple le discours 520 pour lequel ce phénomène se produit. Nous voyons comment la cohérence interne réussit à renverser presque totalement la situation : ainsi, un gros bloc étiqueté *M* par l'adaptation seule, au sein d'un discours dont la classe est *C* a été étiqueté correctement par la cohérence interne, à l'exception de deux phrases dont les probabilités penchaient trop fortement vers la classe *M*.

Le modèle de cohérence interne cherche donc à maximiser la probabilité d'appartenance des phrases proches aux frontières de segments. Il peut utiliser *a)* le réseau de noms propres et *b)* la probabilité issue de l'apprentissage par Markov. Pour un discours S_1^d donné, de longueur *d*, nous cherchons un découpage optimal $\widetilde{D}$ (cf. figure 2) et un étiquetage $\widetilde{E}$ tels que :

$$(\widetilde{D}, \widetilde{E}) = \text{Arg}\max_{D,E} \left\{ P_I(D, E|S_1^d) \times P'(D, E|S_1^d) \right\} \tag{4}$$

où $P'(D, E|S_1^d)$ est la probabilité issue de l'apprentissage et $P_I(D, E|S_1^d)$ la probabilité de cohérence interne (à l'intérieur d'un discours). La conjugaison des modèles d'apprentissage et de cohérence interne est réalisée par le produit entre P' et P_I, qu'il semble légitime de considérer indépendants l'un de l'autre. Comme le découpage ne peut être déduit de l'apprentissage, nous faisons l'hypothèse que $P'(D, E|S_1^d) \cong P'(E|S_1^d)$. Donc :

$$(\widetilde{D}, \widetilde{E}) = \text{Arg}\max_{D,E} \left\{ P_I(D, E|S_1^d) \times P'(E|S_1^d) \right\} \tag{5}$$

Or, d'après le théorème de Bayes :

$$P_I(D, E|S_1^d) = \frac{P(S_1^d|D, E)P(D|E)}{P(S_1^d)} \text{ et } P'(E|S_1^d) = \frac{P'(S_1^d|E)P'(E)}{P'(S_1^d)} \tag{6}$$

De ce fait, l'équation (5) devient :

$$(\widetilde{D}, \widetilde{E}) \cong \text{Arg}\max_{D,E} \left\{ \frac{P(S_1^d|D, E)P(D|E)}{P(S_1^d)} \times \frac{P'(S_1^d|E)P'(E)}{P'(S_1^d)} \right\} \tag{7}$$

Nous savons que $P(D|E)$ prend toujours des valeurs {0, 1} car le découpage est toujours déterminé par les étiquettes (mais pas vice-versa). La probabilité $P'(E)$ ne peut pas être déduite de l'apprentissage (le choix de D peut être considéré comme aléatoire) et $P(S)$ et $P'(S)$ ne dépendent pas de D ou de E. Alors :

$$(\widetilde{D}, \widetilde{E}) \cong \text{Arg}\max_{D,E} \left\{ P_I(S_1^d|D, E) \times P'(S_1^d|E) \right\} \tag{8}$$

Nous avons choisi de représenter un couple (D, E) par un couple de deux indices i et j dont la signification est donnée par la figure 2. Ces deux indices correspondent aux bornes du bloc des segments étiquetés M et à la ligne et la colonne de la matrice Ψ évoquée en section 2.2. Pour un discours donné, on aura donc :

$$\Psi[i, j] = P(S_{1\ldots i-1, j+1\ldots d}|C) \times P(S_{i\ldots j}|M) \times P'(S_{1\ldots i-1, j+1\ldots d}|C) \times P'(S_{i\ldots j}|M) \tag{9}$$

En faisant l'hypothèse[13] que les segments sont indépendants, nous introduisons le produit sur toutes les phrases du discours en distinguant celles qui sont à l'intérieur du bloc $S_{i\rightarrow j}(k =$

[13]Cette hypothèse va quelque peu à l'encontre de l'objectif recherché, à savoir considérer les segments d'un même bloc comme un tout, mais nous ne savons pas comment faire autrement.

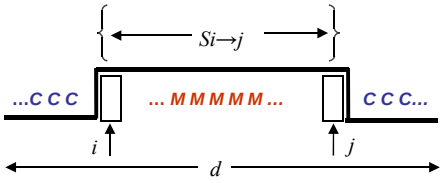

FIG. 2 – *Schéma de découpage des discours.*

$i \ldots j$) de celles qui sont à l'extérieur ($k = 1 \ldots i-1, j+1 \ldots d$) :

$$\Psi[i,j] = \prod_{k=1\ldots i-1, j+1\ldots d} [P(S_k|\chi) \times P'(S_k|C)] \times \prod_{k=i\ldots j} [P(S_k|\mu) \times P'(S_k|M)] \qquad (10)$$

où d est la longueur du discours et $\chi = C \setminus S_k$ et $\mu = M \setminus S_k$. Ceci revient à exclure le segment S_k des données qui servent à estimer les paramètres utilisés pour calculer la probabilité de production de ce même segment S_k. Notons que, si les probabilités $P(S_k|\chi) = 1$ et $P(S_k|\mu) = 1$, alors la valeur de $\Psi[i,j]$ est réduite au cas de Markov (adaptation simple). Nous avons exploité la matrice $\Psi[i,j]$ (cf. figure 3) en nous limitant à sa partie triangulaire supérieure. Le fait d'exclure la diagonale principale dans les calculs illustre l'exploitation de la contrainte respectée par les fournisseurs du corpus DEFT'05. S'il y a des segments de la classe *M* insérés, il y a en au moins deux. Le cas des discours étiquetés uniquement *C* n'est pas représenté dans la figure, mais il a été pris en considération, même s'il n'est pas intégré dans la matrice.

1		...		i	...	d
⋮	•			⋮		
j	•	•		$P(S_i, S_j)$		
	•	•	•	⋮		
⋮	•	•	•	•		
	•	•	•	•	•	⋮
d	•	•	•	•	•	•

FIG. 3 – *Matrice $\Psi[i,j]$ pour le calcul de la cohérence interne. Les • représentent les cases ignorées pour le calcul des probabilités.*

4 Expériences

Pour la Modélisation I, tous les corpus (apprentissage et test) ont été traités par l'ensemble d'outils LIA_TAGG[14]. Ces outils contiennent les modules suivants :

- un module de formatage de texte permettant de découper un texte en unités (ou *tokens*) en accord avec un lexique de référence ;
- un module de segmentation insérant des balises de début et fin de phrase dans un flot de texte, en accord avec un certain nombre d'heuristiques ;
- un étiqueteur morphosyntaxique, basé sur l'étiqueteur ECSTA (Spriet et El-Bèze, 1998) ;
- un module de traitement des mots inconnus permettant d'attribuer une étiquette morphosyntaxique à une forme inconnue du lexique de l'étiqueteur en fonction du suffixe du mot et de son contexte d'occurrence. Ce module est basé sur le système DEVIN présenté dans (Spriet et al., 1996).
- un lemmatiseur associant à chaque couple mot/étiquette morphosyntaxique un lemme en fonction d'un lexique de référence.

Dans la phase de développement, le corpus d'apprentissage a été découpé en cinq sous-corpus de telle sorte que pour chacune des cinq 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 quatre 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 pas de respecter le schéma d'insertion défini dans les spécificités de DEFT'05.

4.1 Résultats de l'adaptation

Des résultats de nos modèles uniquement avec adaptation ont été publiés dans les actes du colloque TALN 2005. Nous reproduisons ici les observations majeures qui pouvaient être faites sur ces résultats. Le *Fscore* s'améliore de façon notable au cours des cinq premières itérations de l'adaptation. 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 (El-Bèze et al., 2005). Nous montrons au tableau 4 et sur la figure 4 les meilleurs *Fscore* officiels soumis pour l'ensemble de participants pour les trois tâches. On peut voir que le système du LIA senior est positionné, dans les trois cas, en première place. La méthode de (Rigouste et al., 2005) en deuxième position, utilise quelques méthodes probabilistes semblables aux nôtres, mais ils partent de l'hypothèse où la segmentation thématique est faite au niveau des orateurs (pas au niveau du discours), ils ont besoin de pondérer empiriquement les noms et les dates (tâches T2 et T3), leurs machines de Markov sont plus complexes et il ne font pas d'adaptation, entre autres.

Le dévoilement des dates (tâche T3) permet d'améliorer très légèrement les résultats du modèle II, mais entraîne une dégradation sur le modèle I. En ce qui concerne la précision et le rappel au fil des itérations sur l'ensemble de Test(T) ainsi que sur le Développement(D), c'est

[14] Téléchargeable à l'adresse : http ://www.lia.univ-avignon.fr

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).

Equipes	T1	T2	T3
1 El-Bèze, Torres, Bechet : LIA senior	**0,87**	**0,88**	**0,88**
2 Rigouste, Cappé, Yvon : ENST	0,86	0,85	0,87
3 Pierron, Durkal, Freydigier : LORIA/UHP	0,82	0,82	0,82
4 Labadie, Romero, Sitbon : LIA junior	0,76	0,74	0,75
5 Maisonnasse, Tambellini : CLIPS	0,75	0,75	0,76
6 Kerloch, Gallinari : LIP6	0,73	0,79	0,79
7 Hernandez : LIMSI	0,56	0,56	0,57
8 Plantié, Dray, Montmain, Meimouni, Poncelet : LGI2P	0,49	0,52	0,51
9 Hurault-Plantet, Jardino, Illouz : LIMSI	0,49	0,56	0,56
10 Chauche : LIRMM	0,32	0,31	0,31
11 Henry, Marley, Amblard, Moot : LABRI	0,18	0,18	0,42

TAB. 4 – *Résultats officiels de l'atelier DEFT'05 du Fscore sur les trois tâches de test {T1, T2, T3} pour les meilleures soumissions de l'ensemble des 11 équipes.*

4.2 Résultats avec la cohérence interne

Les résultats ont été améliorés grâce à la recherche d'une cohérence interne des discours. Cette étape intervient après application de l'automate markovien et avant la phase d'adaptation. Nous montrons, sur les figures 5, 6 et 7, le *Fscore* obtenu pour les trois tâches à l'aide d'une adaptation plus la cohérence interne pour les corpus de Développement (D) et de Test (T). Dans tous les cas, l'axe horizontal représente les itérations de l'adaptation. Sur les graphiques, la ligne pointillée correspond aux valeurs du *Fscore* obtenues avec l'adaptation seule et les lignes continues à celles de la cohérence interne (une itération : ligne grosse ; deux itérations : ligne fine). Pour les trois tâches, on observe une amélioration notable du modèle de cohérence interne par rapport à celui de l'adaptation seule. Enfin, la valeur la plus élevée de *Fscore* est à présent obtenue pour la tâche T3 (figure 7), à un niveau de 0,925. Ce score dépasse largement le meilleur résultat (*Fscore* = 0,88) atteint lors du défi DEFT'05. Les valeurs précises des courbes sont rapportées dans les tableaux 5 et 6.

Les figures 8, 9 et 10 montrent, avec la même convention que les figures précédentes, le *Fscore* pour le modèle II, où n'ont été appliqués ni filtrage ni lemmatisation. Ici encore, les valeurs les plus élevées sont obtenues pour la tâche T3, avec un *Fscore* = 0,873. La comparaison avec les performances (*Fscore* = 0,801 pour la tâche T3) de ce même modèle que nous avons employé lors du défi DEFT'05, est avantageuse : sept points de plus[15]. Cette amélioration est due essentiellement à la cohérence interne et permet d'approcher la meilleure valeur (*Fscore* = 0,881 rapporté dans (El-Bèze et al., 2005)) qui avait été obtenue avec l'adaptation seule et

[15] Plus de détails sont rapportés aux tableaux 7 et 8.

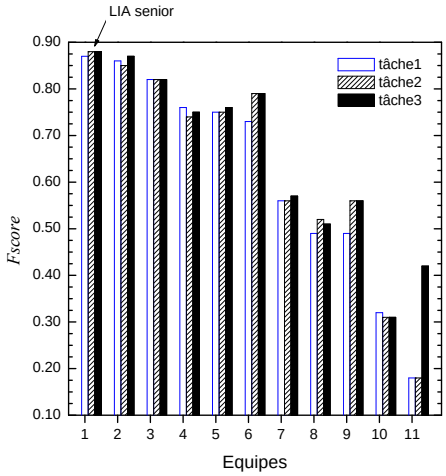

FIG. 4 – *Fscore officiels pour les trois tâches (T1 : pas de noms, pas de dates ; T2 : pas de noms et T3 : avec noms et dates) pour l'ensemble de participants DEFT'05. Les membres des équipes sont cités au tableau 4.*

un filtrage et lemmatisation préalables. Bien que l'utilisation de ce modèle soit un peu moins performante (et de ce fait contestée), nous pensons qu'il peut être utile d'y recourir, si l'on veut éviter la lourdeur des certains processus de prétraitement.

4.3 Fusion de méthodes

Le dernier test que nous avons réalisé fait appel à une fusion de l'ensemble des modèles. Nous avons appliqué un algorithme de vote sur presque toutes les hypothèses issues des modèles I et II. Les hypothèses (qui vont faire office de juges) proviennent des différentes itérations de l'adaptation, avec ou sans cohérence interne. Nous avons tenu compte des avis d'un nombre de juges donné (*NbJ*), en pondérant l'avis de chaque juge *j* par un poids α_j de telle sorte que le critère de décision final est le suivant :

$$\theta_i = signe\left(\sum_{j=1}^{NbJ} \alpha_j \xi_{i,j} - \delta\right) \tag{11}$$

Si θ_i est négatif alors l'étiquette du segment i sera *C* ; *M* autrement. Avec $\alpha_j \in \Re$, $\beta_{i,j} \in \{0,1\}$ et la convention 0 = C et 1 = M.

La stratégie est la suivante : afin d'avoir un degré de confiance suffisant, il faut retenir les segments auxquels une majorité de juges attribue l'étiquette M. Les paramètres α_j et δ ont été ajustés pour minimiser le nombre d'erreurs sur l'ensemble de développement (D). Nous nous sommes proposés de voir cette estimation comme un problème de classification à NbJ entrées et une sortie, c'est-à-dire, comme un problème d'apprentissage supervisé. Nous avons ainsi défini un exemple d'apprentissage comme le vecteur binaire $\xi_j = \{0,1\}, j = 1, \cdots, NbJ$. La

	Tâche T1 (D)		Tâche T2 (D)		Tâche T3 (D)	
It	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation
0	0,841	0,850	0,853	0,852	0,855	0,857
1	0,845	0,871	0,863	0,874	0,864	0,879
2	0,853	0,872	0,863	0,875	0,863	0,880
3	0,857	0,875	0,867	0,877	0,868	0,882
4	0,860	0,877	0,868	0,880	0,868	0,885
5	**0,862**	**0,877**	**0,868**	**0,881**	**0,870**	**0,887**

TAB. 5 – *Modèle I Fscore Développement : Adaptation seule et Cohérence interne.*

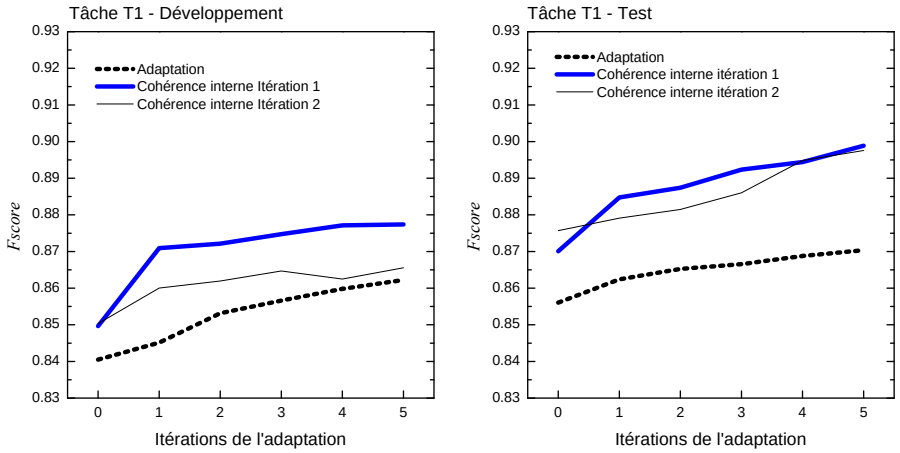

FIG. 5 – *Fscore tâche T1 Modèle I / Adaptation vs Cohérence interne / corpus D et T.*

sortie (classe de référence) de cet exemple est un scalaire $\tau = \{-1, 1\}$ (-1 pour la classe C, $+1$ pour la classe M). L'ensemble d'apprentissage est donc constitué de S segments et NbJ juges, et nous le dénoterons par $\aleph = \{\vec{\xi}_i, \tau_i\}; i = 1, \cdots, S$. Trouver les poids α_j correspond donc à trouver les j poids d'un perceptron entraîné sur l'ensemble $\aleph$. Nous avons utilisé un perceptron optimal à recuit déterministe[16] entraîné par l'algorithme Minimerror (Gordon et Berchier, 1993; Torres Moreno et Gordon, 1995; Torres-Moreno et al., 2002), où l'apprentissage garantit que si l'ensemble $\aleph$ est linéairement séparable, l'algorithme trouve la solution optimale (marge maximale de séparation) et s'il ne l'est pas (comme cela semble être le cas ici), il trouve une solution qui minimise le nombre de fautes commises. Ainsi, nous avons trouvé un seuil δ = 8,834 et les poids α_j avec un nombre de juges NbJ = 89.

Pour l'ensemble de développement, les résultats obtenus au moyen de cette fusion sont encore meilleurs qu'avec les autres méthodes. Nous obtenons, dans ce cas, un *Fscore* = 0,914

[16]La position de l'hyperplan séparateur des classes se fait par une modification progressive des poids (descente en gradient) contrôlés au moyen d'une température de recuit lors de l'apprentissage.

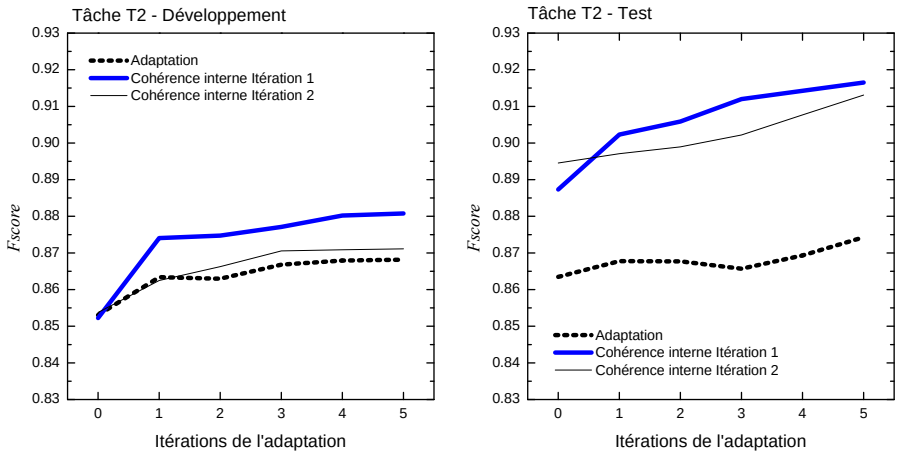

FIG. 6 – *Fscore tâche T2 Modèle I / Adaptation vs Cohérence interne / corpus D et T.*

	Tâche T1 (T)		Tâche T2 (T)		Tâche T3 (T)	
It	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation
0	0,856	0,870	0,863	0,887	0,864	0,887
1	0,862	0,885	0,868	0,902	0,884	0,909
2	0,865	0,888	0,868	0,906	0,884	0,915
3	0,867	0,892	0,866	0,912	0,887	0,920
4	0,869	0,894	0,869	0,914	0,890	0,923
5	**0,870**	**0,899**	**0,874**	**0,917**	**0,897**	**0,925**

TAB. 6 – *Modèle I $Fscore$ Test : Adaptation seule et Cohérence interne.*

avec une précision de 0,916 et un rappel de 0,911. Cependant, pour l'ensemble de test, la fusion ne dépasse pas le meilleur résultat obtenu jusqu'à présent. En effet, on atteint un *Fscore* = 0,914, avec une précision de 0,892 et un rappel de 0,937. Il est connu que les perceptrons (et les réseaux de neurones en général) trouvent parfois des valeurs de poids trop bien adaptées à l'ensemble d'apprentissage (phénomène de sur-apprentissage). Le fait de n'avoir pas eu de meilleurs résultats sur le test le confirme. Cependant, nous pensons que si les $\vec{\xi_i}$ étaient des probabilités au lieu d'être des 0 et des 1, on aurait pu observer un meilleur comportement.

4.4 Analyse des erreurs

Nous avons analysé les erreurs commises par notre système. Sur un total de 27 163 phrases de l'ensemble de Test de la tâche T3, le Modèle I avec la méthode d'adaptation et la cohérence interne des discours, a fait un total de 578 erreurs (*Fscore* = 0,925) :

– 233 erreurs de la classe *C* (faux négatifs assimilés au rappel), dont :

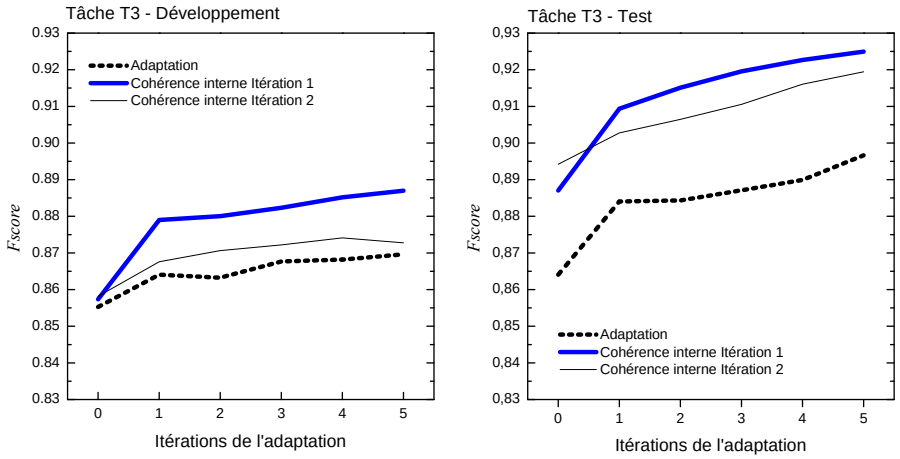

FIG. 7 – *Fscore tâche T3 Modèle I / Adaptation vs Cohérence interne / corpus D et T.*

	Tâche T1 (D)		Tâche T2 (D)		Tâche T3 (D)	
It	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation
0	0,834	0,867	0,842	0,868	0,844	0,867
1	0,844	0,881	0,845	0,881	0,846	0,881
2	0,847	0,881	0,848	0,882	0,849	0,883
3	0,850	0,881	0,851	0,882	0,851	0,883
4	0,854	0,881	0,855	**0,883**	0,854	**0,883**
5	**0,857**	**0,882**	**0,856**	0,882	**0,857**	0,882

TAB. 7 – *Modèle II $Fscore$ Développement : Adaptation seule et Cohérence interne.*

  - 37 phrases *C* à la frontière inversée (≈ 16%) ;
  - 113 phrases *C* en blocs (≈ 49%) ;
  - 83 phrases *C* entre blocs *C* (≈ 36%) ;
- 345 erreurs de la classe *M* (faux positifs assimilés à la précision), dont :
  - 35 phrases *M* à la frontière inversée (≈ 10%) ;
  - 126 phrases *M* en blocs (≈ 37%) ;
  - 184 phrases *M* insérées dans 21 discours de classe *C* exclusive (≈ 53%).

Le problème le plus grave concerne la précision (59% du total des erreurs), et ici, la plus grande majorité (53% de faux positifs) est due aux insertions des phrases *M* dans des discours de classe *C*[17]. L'autre problème se présente dans les 126 phrases en blocs inversés (37%). Ces problèmes sont peut-être dus à l'utilisation de la cohérence interne : sur le tableau 9, on voit qu'en adaptation seule, la précision est toujours plus élevée que le rappel (en D comme

[17] Voir en annexe l'analyse du discours 520, concernant cette situation problématique.

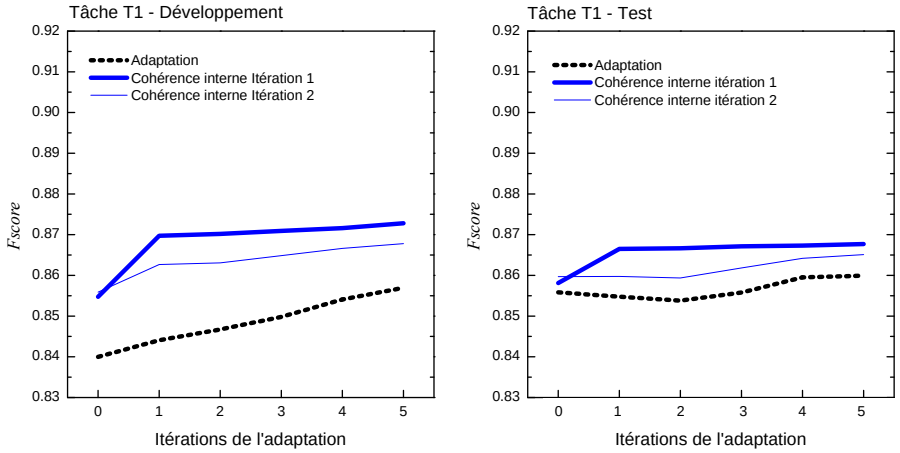

FIG. 8 – *Fscore tâche T1 Modèle II / Adaptation vs Cohérence interne / corpus D et T.*

	Tâche T1 (T)		Tâche T2 (T)		Tâche T3 (T)	
It	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation	Adaptation seule	Cohérence puis adaptation
0	0,856	0,857	0,853	0,860	0,852	0,859
1	0,855	0,873	0,853	0,877	0,852	0,872
2	0,854	0,873	0,853	0,877	0,853	0,873
3	0,856	**0,873**	0,855	0,877	0,854	0,874
4	0,860	0,872	**0,859**	0,877	0,856	0,874
5	**0,860**	0,871	0,858	**0,877**	**0,856**	**0,874**

TAB. 8 – *Modèle II $Fscore$ Test : Adaptation seule et Cohérence interne.*

en T). Pour la cohérence interne, la situation est inversée : le rappel est bien meilleur que la précision. Le même comportement a été retrouvé dans le Modèle II. Un autre pourcentage important d'erreurs (49% de faux positifs) a lieu dans l'inversion d'un nombre important de blocs (113 phrases). Enfin, une autre partie non négligeable (10% de faux positifs, 16% de faux négatifs) correspond à l'inversion de catégorie d'une phrase unique à la frontière des découpages (soit i ou j, voir figure 2). La détection de cette frontière, reste un sujet très délicat avec nos approches.

La figure 11 montre les courbes de Précision et de Rappel. Pour rester concis, nous allons présenter ici seulement les résultats correspondant à la tâche T3 du modèle I. La cohérence interne est affichée uniquement sur la première itération. Dans les deux cas, nous montrons des résultats sur les corpus de Développement (D) et Test (T). On voit que sur l'ensemble de test T, la précision de la cohérence interne puis adaptation est moins élevée que celle de l'adaptation seule. La même situation se produit pour le rappel. Néanmoins, un phénomène d'inversion se présente en développement : en rappel on est plus performant avec la cohérence qu'avec

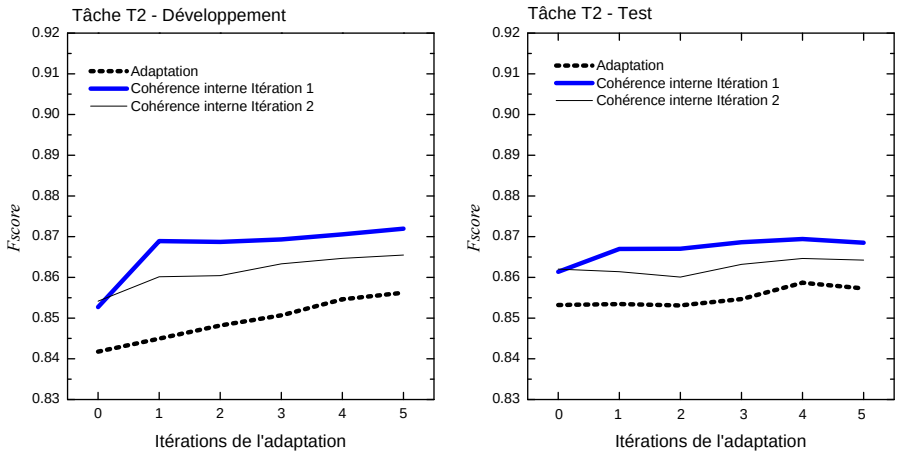

FIG. 9 – *Fscore tâche T2 Modèle II / Adaptation vs Cohérence interne / corpus D et T.*

	Adaptation		Cohérence interne puis adaptation	
	Précision	Rappel	Précision	Rappel
Développement	**0,918**	0,826	0,882	**0,892**
Test	**0,931**	0,866	0,912	**0,939**

TAB. 9 – *Précision et Rappel pour la tâche T3, Modèle I, adaptation et cohérence interne à la dernière itération de l'adaptation.*

l'adaptation seule, et en précision avec l'adaptation seule que avec la cohérence. Nous avons aussi calculé la longueur moyenne (en mots) des segments mal classés (tâche T3 / modèle I) suivant le type d'erreur (cf. figure 12) : discours exclusifs de la classe *C*, longueur moyenne ≈ 21. Erreurs de début du bloc : type 1 longueur moyenne ≈ 21 ; type 2 longueur moyenne ≈ 22. Erreurs de fin du bloc : type 3 longueur moyenne ≈ 21 ; type 4 longueur moyenne ≈ 26. Cependant, il est difficile de tirer de conclusions à partir de cette information. Enfin, de façon comparative les résultats du *Fscore* sur les trois tâches sans adaptation, montrés au tableau 10, confirment l'importance de l'utilisation de l'automate de Markov : il fait un gain global de $\approx 25\%$ avec les deux modèles proposés.

	Développement			Test		
	T1	T2	T3	T1	T2	T3
Modèle I	0,570	0,570	**0,574**	0,593	0,595	**0,596**
Modèle II	0,549	0,551	**0,555**	0,581	0,582	**0,585**

TAB. 10 – *Fscore pour les trois tâches sans adaptation.*

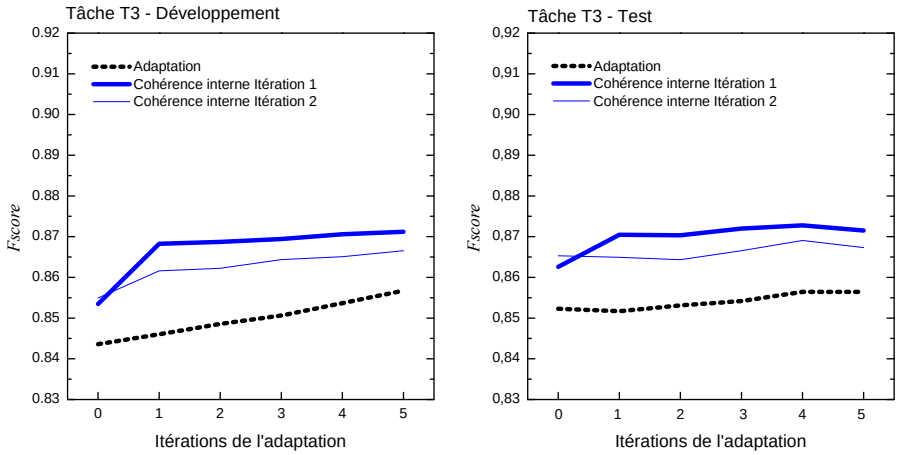

FIG. 10 – *Fscore tâche T3 Modèle II / Adaptation vs Cohérence interne / corpus D et T.*

5 Conclusions et perspectives

Nous avons introduit une formalisation de la cohérence interne des discours qui a beaucoup amélioré nos résultats rapportés en (El-Bèze et al., 2005). Cette cohérence ainsi que l'adaptation ont été combinées conjointement avec les modèles d'apprentissage préalablement développés, comme la modélisation bayésienne qui semble déterminant, l'automate de Markov et des processus d'adaptation. Les résultats que nous obtenons pour la tâche T3 avec la cohérence interne en terme de *Fscore* = 0,925 sont très encourageants. Cependant, l'utilisation de la cohérence interne présente un risque : quelques phrases avec une thématique différente, peuvent faire basculer tout un bloc vers l'autre étiquette. Ce type de comportement local entraîne des instabilités globales (semblables à ce qui se produit dans le jeu « Reversi »), dont la prévision reste très difficile, ayant comme conséquence une baisse générale des performances. Ne pas lemmatiser et ne rien filtrer dégrade un peu les performances (*Fscore* = 0,874 avec le modèle II) mais permet d'éviter l'application d'un processus additionnel de prétraitement qui pour certaines langues peut être relativement lourd. La fusion des hypothèses vue comme un vote de plusieurs juges pondérés par un perceptron optimal a permis de surpasser les résultats précédents en développement (*Fscore* = 0,914). Cependant nous pensons qu'il reste encore du travail pour améliorer cette stratégie afin d'obtenir de meilleures performances en test. Des études comme celle de (Rigouste et al., 2005) sur le même corpus confirment que l'utilisation de méthodes probabilistes est la mieux adaptée à ce type de segmentation thématique. Le recours à un réseau de Noms Propres est utile et nous encourage par la suite à employer une ressource lexicale comme (Vossen, 1998) pour tirer parti de réseaux sur les noms communs. Pour s'affranchir des contraintes liées à la constitution d'une ressource sémantique, il serait judicieux de recourir à des approches telles que *Latent Semantic Analysis* (Deerwester et al., 1990) ou PLSA (Hofmann, 1999). D'autres perspectives d'application, comme celle de la séparation de thèmes sont aussi envisageables. Il faut reconnaître, cependant, que s'il s'était agi de traiter un texte composite moins artificiel que celui proposé par DEFT, par exemple un dia-

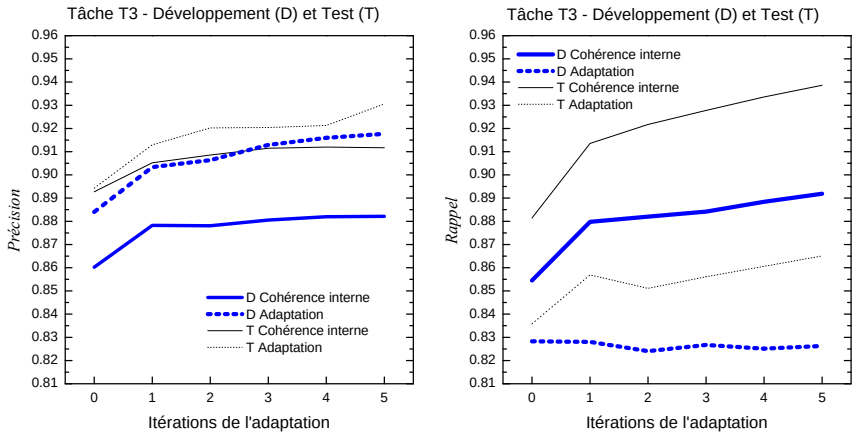

FIG. 11 – *Précision-Rappel Modèle I / tâche T3 : Adaptation vs Cohérence interne.*

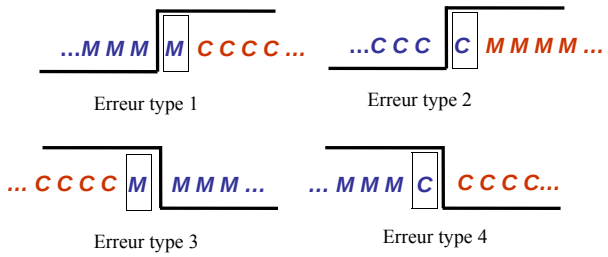

FIG. 12 – *Types d'erreurs frontière de bloc.*

logue, la difficulté aurait été accrue. 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 orateurs ou plusieurs thèmes comme envisagé dans le cadre du Projet Carmel (Chen et al., 2005).

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 *Proc. of TALN 2005 - Atelier DEFT'05*, Volume 2, pp. 95–97.

Azé, J. et M. Roche (2005). Présentation de l'atelier DEFT'05. In *Proc. of TALN 2005 - Atelier DEFT'05*, Volume 2, pp. 99–111.

Bellot, P., E. Crestan, M. El-Bèze, L. Gillard, et C. D. Loupy (2003). Coupling named entity recognition, vector-space model and knowledge bases for TREC-11, question-answering track. In *Proceedings of TREC'02, Gaithersburg, USA, NIST Special publication 500 251*.

Chen, B., M. El-Bèze, M. Haddara, O. Kraif, et G. M. de Montcheuil (2005). Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale. In *Proc. of Traitement Automatique des Langues Naturelles 2005*, Volume 1, pp. 415–418.

Collobert, R. et S. Bengio (2001). Support vector machines for large-scale regression problems. *Journal of Machine Learning Research 1*, 143–160.

Cotteret, J.-M. et R. Moreau (1969). *Le vocabulaire du Général de Gaulle*. Armand Colin, Presses de la Fondation nationale des sciences politiques.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman (1990). Indexing by latent semantic analysis. *American Society For Information Science 41*, 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 *Proc. of TALN 2005 - DEFT'05*, Volume 2, pp. 125–134.

Freund, Y. et R. Schapire (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences 55*, 119–139.

Gordon, M. et D. Berchier (1993). Minimerror: A perceptron learning rule that finds the optimal weights. In M. Verleysen (Ed.), *ESANN*, Brussels, pp. 105–110. D facto.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proc. of 2nd Annual ACM Conference on Research and Development in Information Retrieval, Berkeley, California*, pp. 50–57.

Illouz, G., B. Habert, S. Fleury, H. Folch, S. Heiden, P. Lafon, et S. Prévost (2000). Profilage de textes : cadre de travail et expérience. In *JADT 2000, Journées Int. d'Analyse Statistiques des Données Textuelles, Lausanne*, pp. 163–170.

Jalam, R. et J.-H. Chauchat (2002). Pourquoi les n-grammes permettent de classer des textes ? recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques. In *JADT 2002, Journées Int. d'Analyse Statistiques des Données Textuelles, St-Malo*, pp. 13–15.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing 35*, 400–401.

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

Manning, C. D. et H. Schütze (2000). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Rigouste, L., C. Olivier, et Y. François (2005). Modèle de mélange multi-thématique pour la fouille de textes. In *Proc. of TALN 2005 - Atelier DEFT'05*, Volume 2, pp. 193–202.

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

Schapire, R. et Y. Singer (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning 39(2/3)*, 135–168.

Soboro, I. (2004). Overview of the TREC 2004 Novelty Track. In E. M. Voorhees et L. P. Buckland (Eds.), *Proceedings of TREC'04*, USA. NIST Special Publication: SP.

Spriet, T., F. Béchet, M. El-Bèze, C. D. Loupy, et L. Khouri. (1996). Traitement automatique des mots inconnus. In *Proc. of TALN 96*, Marseille France, 22-24 mai, pp. 170–179.

Spriet, T. et M. El-Bèze (1998). Introduction of Rules into a Stochastic Approach for Language Modelling. *Computational Models of Speech Pattern Processing 169*, 350–355.

Torres-Moreno, J.-M., J. Aguilar, et M. Gordon (2002). Finding the number minimum of errors in N-dimensional parity problem with a linear perceptron. *Neural Processing Letters 1*, 201–210.

Torres Moreno, J.-M. et M. Gordon (1995). An evolutive architecture coupled with optimal perceptron learning. In M. Verleysen (Ed.), *ESANN*, Brussels, pp. 365–370. D facto.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic publishers.

Annexe

Nous souhaitons illustrer le fonctionnement du modèle de cohérence interne en rapportant ici une étude sur le discours 520 du corpus de test, dont toutes les phrases appartiennent à la classe C (Chirac). Le tableau 11 montre des extraits des phrases de ce discours, ainsi que leurs probabilités $p(C)$ d'appartenance à la classe C (probabilité $p(M) = 1 - p(C)$) calculées par adaptation et celles p_I obtenues avec la cohérence interne du discours (cf. section 3), puis adaptation. L'étiquette associée est aussi indiquée. En gras, nous affichons les mots pleins (**Afrique, africains, Congo, développ(er/ment)**,...) des phrases étiquetées C. En souligné les mots pleins communs (effort, institutions) aux deux classes, et en petites majuscules les mots pleins (ACCORD) présents uniquement à l'intérieur du bloc M. Les adjectifs, adverbes, pronoms, conjonctions de subordination ainsi que les noms propres trop fréquents dans le corpus DEFT'05 (tels que *France, français(e), Paris, international, pays,...*) ont été supprimés pour ne pas fausser les calculs. Les phrases 39 et 40 (en italiques) méritent d'être analysées en détail. Leurs probabilités d'appartenance à la classe M, calculées uniquement avec l'adaptation seule, étaient déjà très élevées : 0,671 ($p(C) = 0,329$) et 0,853 ($p(C) = 0,147$) respectivement (et sont d'ailleurs les plus élevées de tout ce discours), donc difficiles à renverser. Elles faisaient partie d'un gros bloc (phrases 31-40) étiqueté M par Markov. Après le calcul de la cohérence, les phrases 39 et 40 seront encore étiquetées comme M, avec des valeurs 0,91 ($p(C) = 0,09$) et 0,89 ($p(C) = 0,11$) respectivement. Cependant, la méthode de la cohérence interne a fait basculer vers la classe C sept phrases (31-38) du bloc original, étiquetées M à tort dans un premier temps. Ce renversement n'est en rien négligeable comme en atteste l'augmentation du Fscore due au mécanisme de la cohérence. La figure 13 montre les probabilités $p(C)$ en fonction du numéro de la phrase du discours. En pointillé, nous affichons les probabilités de l'adaptation seule et en gras continu celles de la cohérence interne, puis adaptation. On voit que, sauf pour les deux phrases en question, toutes les autres ont été réarrangées de façon à avoir des probabilités $p(C)$ bien au-dessus de leurs valeurs précédentes (plusieurs phrases ont maintenant des probabilités $p(C) = 1$). La rupture de cohérence interne du discours reflète ici plutôt un changement thématique : tout le discours 520 s'inscrit fortement dans une vision politique centrée sur l'Afrique, alors que les phrases 39 et 40 parlent soudain de politique

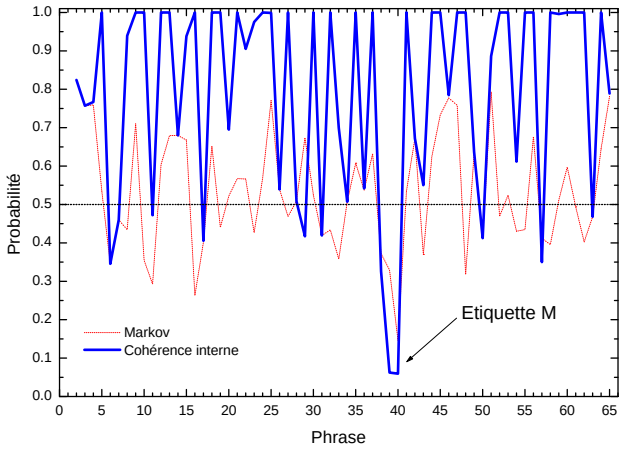

FIG. 13 – *Probabilités $p(C)$ de Markov (ligne pointillée) et de la cohérence interne puis adaptation (trait gras) du discours 520 (dont tous les segments appartiennent à la classe C).*

internationale dans un sens beaucoup plus large (*New York, Washington, institutions internationales*). Ceci illustre la tendance (toujours difficile à modéliser) qu'a parfois un locuteur d'introduire subitement des changements (quasi aléatoires) dans son discours. Cet honneur revient dans le cas présent, à Jacques Chirac.

Remerciements

Nous remercions Eric Gaussier du Centre de Recherche Xerox Grenoble d'avoir mis à notre disposition un lexique de Noms Propres. Nous sommes également reconnaissants envers Jérome Azé et Mathieu Roche du LRI pour l'organisation de DEFT'05.

Summary

We present a set of probabilistic models applied to binary classification as defined in the DEFT'05 challenge. The challenge consisted a mixture of two differents problems in Natural Language Processing : identification of author (a sequence of François Mitterrand's sentences might have been inserted into a speech of Jacques Chirac) and thematic break detection (the subjects addressed by the two authors are supposed to be different). Markov chains, Bayes models and an adaptative process have been used to identify the paternity of these sequences. A probabilistic model of the internal coherence of speeches which has been employed to identify thematic breaks. Adding this model has shown to improve the quality results. A comparison with different approaches demostrates the superiority of a strategy tha combines learning, coherence and adaptation. Applied to the DEFT'05 data test the results in terms of precision (0.890), recall (0.955) and *Fscore* (0.925) measure are very promising.

Discours 520 (Test) Phrase	Probabilités C p	p_I
...13 Je dois donc saluer et remercier tout particulièrement l'ensemble d'entre vous qui, civils, militaires, religieux, coopérants, industriels, professionnels libéraux, artisans, commerçants, etc... sont la France, la France au **Congo**.	0,68 C	1,00 C
...21 On voit qu'un peu partout, et notamment au **Congo**, l'état de droit, la démocratie, est en train de prendre racine et de se **développer**.	0,57 C	1,00 C
...24 De la même façon, on voit les efforts considérables qui sont engagés en **Afrique**, en **Afrique** francophone, au **Congo**, pour libéraliser l'économie, sortir des structures paralysantes, qui, longtemps, ont caractérisé beaucoup de pays **africains** pour donner une nouvelle impulsion à l'initiative, pour gérer avec plus de soin, de sérieux et de transparence, les fonds publics.	0,57 C	1,00 C
...27 On connaît encore des problèmes, des crises, ici ou là, pas loin du **Congo**, dans cette **Afrique**.	0,47 C	1,00 C
...29 On voit des efforts de coordination régionale qui sont engagés et qui permettront une plus grande synergie de l'effort économique et donc du progrès.	0,67 C	0,42 C
30 Je le disais à Franceville tout à l'heure, il y a cinq ans, il y avait peine 20 pays qui, en **Afrique**, avaient une croissance positive par tête d'habitant.	0,53 C	1,00 C
31 Il y en a 41 aujourd'hui en quelques années.	0,42 M	0,51 C
32 Celles ou ceux qui sont **afro** pessimistes sont nombreux, notamment dans le reste du monde, (certains parce qu'ils sont tout simplement découragés, certains parce qu'ils sont ignorants ou veulent l'être de ce qui se passe ici, certains parce qu'ils veulent en fait se désengager de l'aide que le monde industrialisé doit au titre de la solidarité à l'**Afrique** en **développement**).	0,43 M	1,00 C
33 Il y a une espèce de culte d'**afro** pessimisme, il faut aujourd'hui comprendre qu'il n'y a plus aucune raison de **développer** ce sentiment.	0,36 M	0,78 C
34 Et que l'on peut aujourd'hui, raisonnablement, justement, être **afro** optimiste.	0,51 M	0,58 C
35 La croissance du **Congo** est de 6%.	0,61 M	1,00 C
36 J'imagine la satisfaction des Français s'ils faisaient la même performance, nous n'aurions plus aucun problème.	0,54 M	0,99 C
37 Il faut encourager les **Africains**.	0,63 M	1,00 C
38 On ne dit pas assez les efforts considérables, parfois avec des maladresses dues souvent à l'inexpérience, qu'ils ont fait pour **redresser** la situation.	0,37 M	0,39 C
39 Il y a quelques années, seuls quelques pays avaient un ACCORD *avec les institutions internationales, aujourd'hui presque tous sont dans ce cas.*	***0,33 M***	***0,09 M***
40 Je sais bien qu'il est de bon ton, je l'ai fait moi-même souvent, de critiquer des institutions internationales qui, depuis New York ou Washington, depuis les bureaux climatisés et qui sont là-bas à partir des ordinateurs qui s'y trouvent, imposent des règles, non seulement extrêmement difficiles à accepter dans les pays qui doivent faire un effort d'ajustement structurel, mais de plus, le font souvent dans des termes qui ne sont même pas compris, ici, là où leurs règles doivent s'appliquer.	***0,15 M***	***0,11 M***
41 Mais il faut dire aussi, que le temps passant, il y a une amélioration sensible de l'approche, de la vision portée par ces institutions sur l'**Afrique**.	0,54 C	1,00 C
...43 Alors des progrès sont accomplis, il faut naturellement tout faire pour les **développer**.	0,37 M	0,64 C
44 La France, vous le savez, est très attachée à sa politique **africaine**.	0,63 M	1,00 C
45 Elle l'est bien sûr en raison des liens très anciens qui nous unissent, l'**Afrique** ou une partie de l'**Afrique** et nous-mêmes.	0,73 M	1,00 C
...47 C'est en **Afrique** qu'elle a puisé l'énergie, le courage, la détermination, le sang qui nous a permis de **redresser** notre situation si compromise.	0,76 M	1,00 C
...57 Il y a eu, c'est, vrai, une décennie mauvaise, certains ont dit une décennie perdue, et bien, nous nous en sommes sortis, et donc maintenant nous devons faire un effort.	0,41 M	0,41 C
...64 Et bien, je voudrais vous donner ce soir, c'est mon dernier mot, un message de confiance et d'espoir, d'encouragement aussi, et vous dire que la France est fière d'avoir ses meilleurs enfants sur cette terre d'**Afrique**, au **Congo** ou ailleurs, et vous dire que votre rôle est capital pour cette terre et aussi pour les valeurs, les valeurs morales qui sont les nôtres et que j'évoquais il y a un instant...	0,65 M	1,00 C

TAB. 11 – *Exemple de découpage de la cohérence interne. Discours 520 du Test.*