Identification des auteurs de textes courts avec des n-grammes de caractères

Michèle Jardino

LIMSI – CNRS – 91403 Orsay – France

Abstract

A binary authorship attribution based on character level n-gram language models is presented. This simple method has been applied on speeches which were segmented in short sentences. The goal was to detect Mitterrand sentences among Chirac sentences in the frame of the french DEFT'05 evaluation. Our best results are obtained with character-based 4-gram models smoothed with the Viterbi algorithm. They are comparable with more sophisticated works which were presented in the frame of this competition.

Résumé

Nous montrons que des modèles de langage appris à partir des fréquences de suites de n caractères calculées sur des textes d'auteurs identifiés permettent de reconnaître les auteurs de textes courts non signés quasiment aussi bien que des modèles fondés sur les mots ou les analyses syntaxiques des phrases. Ceci a été évalué et comparé aux résultats de l'évaluation DEFT'05 sur la détection de phrases de Mitterrand dans des discours de Chirac. Les meilleurs résultats obtenus avec des quadrigrammes de caractères (F=55%) sont très améliorés après un lissage avec l'algorithme de Viterbi (F=78%). Ils arrivent au 4ème rang dans l'évaluation DEFT'05 derrière des modèles plus complexes mais pour la plupart également lissés avec l'algorithme de Viterbi.

Mots-clés : identification d'auteur, signature de textes, modèle de langage statistique, n-grammes de caractères, algorithme de Viterbi

1. Introduction

Après avoir participé (Hurault-Plantet et al, 2005) au DEfi Fouille de Textes, DEFT'05, (Alphonse et al, 2005) qui consistait à reconnaître des phrases de Mitterrand incluses dans des discours de Chirac et examiné l'ensemble des méthodes utilisées par les participants, il nous est apparu qu'un des traits caractéristiques des textes à savoir le caractère, n'avait pas été utilisé. Or ce trait paraît particulièrement adapté à l'identification de phrases courtes par des méthodes statistiques car sa redondance y est plus importante que celle du mot ou des entités construites sur le mot (lemmes, parties du discours …).

Dès le début du 20ème siècle, Markov (Markov, 1913) calculait les probabilités qu'une voyelle succède à une consonne ou à une autre voyelle à partir des fréquences des suites de deux lettres (bigrammes de lettres) observées dans un poème de Pouchkine. Il montrait des différences significatives entre ces deux probabilités et celles mesurées sur une nouvelle d'Aksakov, un autre écrivain. En 1951, Shannon (Shannon, 1951) calculait l'entropie de l'anglais à partir des prédictions faites par des humains sur des successions de lettres (n-grammes de lettres). Aujourd'hui, la redondance des n-grammes de caractères est largement utilisée en compression des textes électroniques et pour l'identification des langues (Juola, 1998) et de manière plus secondaire en reconnaissance d'auteurs (Kmelev et Tweeddie, 2001 ; Teahan, 2000). Le point important est d'adapter tailles d'apprentissage et de

test à la tâche (Peng et al, 2003). Ces derniers auteurs mentionnent des taux de reconnaissance de 98% pour 8 auteurs au style très différent avec des modèles 6-grammes de caractères appris sur des corpus de grande taille, de l'ordre du million de caractères.

Notre problème se distingue par le fait qu'il s'agit d'identifier au sein de textes écrits par un auteur, des textes courts, d'une centaine de caractères écrits par un autre auteur dans la même langue.

2. Corpus d'apprentissage de DEFT'05

Le corpus fourni par les organisateurs de DEFT'05 se compose de 587 discours de Chirac dont 400 contiennent une séquence de phrases de Mitterrand. Trois formats différents de ce corpus ont été proposés. Comme nous n'avons pas observé de différence significative dans les résultats obtenus sur ces trois formats nous ne présentons ici que le premier format pour lequel les noms de personnes et années ont été supprimés et remplacés par des balises <nom> et <date>. Nous avons formaté ce corpus en remplaçant toutes les majuscules par des minuscules et en conservant tous les signes de ponctuation. Nous avons ensuite séparé tous les caractères. Quelques statistiques sur ce corpus sont rassemblées dans le tableau 1.

Auteur	Nb phrases	Nb mots			Nb caractères		
		Lexique	total	moyenne/ phrase	Lexique	total	moyenne/ phrase
Chirac	49 890	25 534	1 250 901	25 (16)	80	5 504 100	110 (69)
Mitterrand	7 523	13 961	246 552	33 (24)	64	1 024 536	136 (100)

Tableau 1 : Statistiques du corpus d'apprentissage, entre parenthèses sont indiquées les écarts-type

Les caractères comportent donc les lettres en minuscule (incluant les lettres accentuées même peu fréquentes), les chiffres de 0 à 9 et les signes de ponctuation. Les différences de taille entre les lexiques de caractères de Chirac et Mitterrand proviennent de lettres accentuées peu fréquentes. Nous avons préféré garder ces informations pour minimiser les interventions sur le corpus initial.

Le tableau 1 montre une redondance moyenne des caractères 1000 fois supérieure à celle des mots aussi bien pour Chirac que pour Mitterrand ce qui devrait entraîner une meilleure fiabilité des modèles fondés sur les caractères comparée à celle des modèles fondés sur les mots.

Les phrases de Chirac sont en moyenne plus courtes (25 mots) que les phrases de Mitterrand (33 mots). La longueur des phrases a été utilisée comme indicateur par un des participants de DEFT'05 (Pierron, 2005) pour tenter de mieux déceler les blocs contigus de phrases de Mitterrand mais sans résultats probants.

Le nombre moyen de caractères par mot est de 4 avec un écart-type de 3 aussi bien pour Chirac que pour Mitterrand.

2.1. n-grammes de caractères

Les n-grammes de caractères ont été calculés avec les logiciels de CMU (Clarkson, 1997) sur l'ensemble des phrases de Chirac et sur l'ensemble des phrases de Mitterrand. Des modèles de langage probabilistes sont créés à partir de ces n-grammes. Ils prédisent un caractère connaissant les n-1 caractères précédents (modèles de Markov d'ordre n-1). Les probabilités des évènements non observés à l'apprentissage sont calculées par une méthode de repli sur les

probabilités des n-grammes d'ordre inférieur avec une pondération qui prend en compte le nombre de contextes dans lesquels sont vus les n-grammes (Witten et Bell, 1991). Pour chaque valeur de n nous disposons de deux modèles, un par auteur, qui nous permettent de calculer deux valeurs de probabilités pour chaque phrase. En représentant une phrase de longueur L par la succession de ses caractères : $c_1 \ldots c_i \ldots c_L$, la probabilité de cette phrase P_A^C *(phrase)*, est calculée à partir du modèle n-grammes pour l'auteur A :

$$P_A^C(phrase) = \prod_{i=1}^{i=L} p_A(c_i / c_{i-n+1} \ldots c_{i-1})$$

où $p_A\ (\ c_i\ /\ c_{i-n+1} \ldots c_{i-1}\ )$ est la probabilité que le caractère c_i succède à la chaîne de n-1 caractères $c_{i-n+1} \ldots c_{i-1}$. Cette probabilité est calculée à partir des fréquences relatives de la chaîne $c_{i-n+1} \ldots c_{i-1}\ c_i$ dans les phrases de l'Auteur. De cette valeur on déduit une probabilité moyenne de la phrase qui est la racine L$^{\text{ième}}$ de P_A^C, soit la moyenne géométrique de P_A^C sur la longueur L de la phrase.

L'auteur d'une phrase non signée est celui dont le modèle confère la plus grande probabilité moyenne à la phrase.

3. Évaluation

Nous avons repris les corpus de test de DEFT'05 que nous avons formaté de la même manière que les textes d'apprentissage. Dans le tableau 2 suivant sont répertoriées quelques statistiques de ce corpus qui correspond également aux textes dans lesquels noms de personnes et années ont été transformés. Les nombres moyens de mots et de caractères par phrase sont comparables à ceux du corpus d'apprentissage.

Nb discours	Nb phrases	Nb mots	Nb carractères	Nb mots moyen/phrase	Nb caractères moyen/phrase
294	27 162	701 427	3 060 230	26 (17)	113 (72)

Tableau 2 : Statistiques du corpus de test

3.1. Résultats

Nous avons identifié les phrases de Mitterrand dans les discours de Chirac avec des modèles n-grammes de caractères. Pour comparaison nous avons également utilisé des modèles n-grammes de mots comme nous l'avions fait pour le défi initial (Hurault-Plantet et al, 2005). Nous avons traité tous ces résultats à l'aide d'un algorithme de Viterbi pour prendre en compte le fait qu'une seule séquence de phrases de Mitterrand pouvait être incluse dans un discours de Chirac.

3.1.1. N-grammes de caractères

À partir du corpus d'apprentissage nous avons créé des modèles n-grammes de caractères séparés pour Chirac et Mitterrand, pour des valeurs de n allant de 1 à 6. Puis nous avons calculé les probabilités moyennes de chaque phrase du corpus de test avec les modèles Chirac et Mitterrand. Pour chaque n, la phrase est attribuée à l'auteur qui donne la plus grande probabilité moyenne.

Ensuite nous avons comparé ces résultats avec la référence, ce qui nous a permis de calculer des valeurs de rappel R, précision P et F-mesure pour l'identification des phrases de

Mitterrand. Si M est le nombre de phrases de Mitterrand dans le corpus de test, m, le nombre de phrases attribuées à Mitterrand par notre système, et M_m le nombre de phrases de Mitterrand dans m, on a :

$$\text{Précision} = M_m / m$$

$$\text{Rappel} = M_m / M$$

$$\text{F-mesure} = 2 * \text{Précision} * \text{Rappel} / (\text{Précision} + \text{Rappel})$$

Les valeurs obtenues sont répertoriées dans le tableau 3 suivant et écrites en italique :

n	1	2	3	4	5	6
Rappel	*0,54* 0,48	*0,66* 0,72	*0,70* 0,78	*0,65* 0,71	*0,53* 0,52	*0,44* 0,38
Précision	*0,24* 0,52	*0,31* 0,62	*0,37* 0,74	*0,47* 0,87	*0,56* 0,91	*0,60* 0,91
F-mesure	*0,34* 0,50	*0,42* 0,66	*0,49* **0,75**	***0,55*** **0,78**	***0,55*** 0,66	*0,51* 0,54

Tableau 3 : Identification des phrases de Mitterrand dans des discours de Chirac en termes de Rappel, Précision et F-mesure pour différents n-grammes de caractères. La deuxième ligne de chaque rangée correspond à ces mêmes modèles complétés par un lissage de Viterbi.

Les meilleures valeurs sont obtenues pour n = 4 et 5.

Dans l'évaluation DEFT'05 (Alphonse et al, 2005), plusieurs participants ont utilisé avec succès l'algorithme de Viterbi pour lisser leurs résultats en tenant compte des contraintes de la tâche. Nous avons donc implémenté cet algorithme en imposant zéro ou une seule insertion d'une séquence d'au moins deux phrases de Mitterrand dans chaque discours de Chirac. Nous avons ainsi déterminé un schéma de transition entre phrases d'un discours avec quatre états possibles : 2 états Mitterrand M1 et M2 pour prendre en compte le fait qu'au moins 2 phrases de Mitterrand sont incluses dans chaque allocution, et 2 états C1 et C2 pour les phrases de Chirac, un pour les phrases du début d'allocution et un pour les phrases de fin d'allocution pour tenir compte de l'inclusion des phrases de Mitterrand.

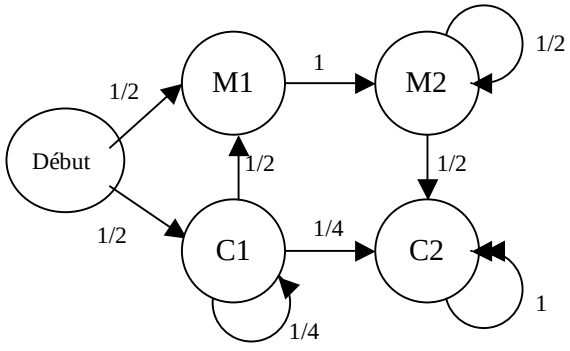

Figure 1 : Probabilités de transition entre les quatre états possibles des phrases d'une allocution, M1 et M2 pour les phrases de Mitterrand et C1 et C2 pour les phrases de Chirac.

Les résultats initiaux sont spectaculairement augmentés, de 0,55 à 0,78 pour n = 4. Ils s'intercalent au 4ème rang de l'évaluation DEFT'05. On peut remarquer que la valeur de n = 4 correspond à la longueur moyenne des mots.

Les expériences précédentes ont été faites avec et sans le caractère « ESPACE » avec les mêmes résultats pour les valeurs de n jusqu'à 4. On constate seulement une décroissance plus lente des performances lorsque l'on introduit le caractère « ESPACE » quand n augmente au-delà de 4. Il est à noter que 98% des 4-grammes du test ont été observés dans le corpus d'apprentissage.

3.1.2. N-grammes de mots

Pour rappel nous avons repris les mêmes expériences avec des n-grammes de mots, les résultats sont présentés dans le tableau 4 suivant.

n	1	2	3	4	6	8
Rappel	*0,83* 0,91	*0,62* 0,65	0,58 0,60	*0,59* 0,60	*0,60* 0,61	*0,60* 0 ,61
Précision	*0,21* 0,29	*0,36* 0,72	0,40 0,79	*0,41* 0,79	*0,40* 0,78	*0,39* 0,74
F-mesure	*0,34* 0,44	*0,46* **0,68**	***0,48*** **0,68**	***0,48*** **0,68**	***0,48*** **0,68**	*0,47* 0,67

Tableau 4 : Identification des phrases de Mitterrand dans des discours de Chirac en termes de Rappel, Précision et F-mesure pour différents n-grammes de mots. La deuxième ligne de chaque rangée correspond à ces mêmes modèles complétés par un lissage de Viterbi.

Le maximum de performances est plat de n=3 à n=6. On remarque comme pour les modèles n-grammes de caractères, que les performances sont notablement accrues avec le lissage de Viterbi. Elles restent néanmoins inférieures de 10 points aux performances des modèle n-grammes de caractères lissés.

3.2. Pouvoir discriminant des modèles

Plusieurs travaux mettent en évidence une entropie moyenne caractéristique d'un auteur. Nous avons voulu mesurer les distributions de l' « entropie » des phrases courtes des corpus avec les meilleurs modèles, ceux créés à partir des 4-grammes de caractères. Nous voulons ainsi vérifier si l'utilisation de caractères a un pouvoir discriminant suffisant, ou en d'autres termes si les n-grammes de caractères peuvent constituer une signature d'auteur. Il existe un lien direct entre la probabilité d'une phrase P_A^C *(phrase)* et son entropie H_A^C *(phrase)* (Cover, 1991) qui est :

$$H_A^C \text{ (phrase)} = -(1/L) \log (P_A^C \text{ (phrase)})$$

H_A^C *(phrase)* est une grandeur moyenne qui permet de s'affranchir de la longueur des phrases et qui varie entre 0 et la taille du lexique, en sens inverse de la probabilité de la phrase.

Nous avons relevé pour chaque phrase du corpus de test les valeurs d'entropie données par les deux modèles 4-grammes de caractères de Chirac et Mitterrand et compté combien de phrases de Chirac puis de Mitterrand avaient la même entropie avec une précision de l'ordre du %. Ces valeurs sont représentées sur la figure 2.

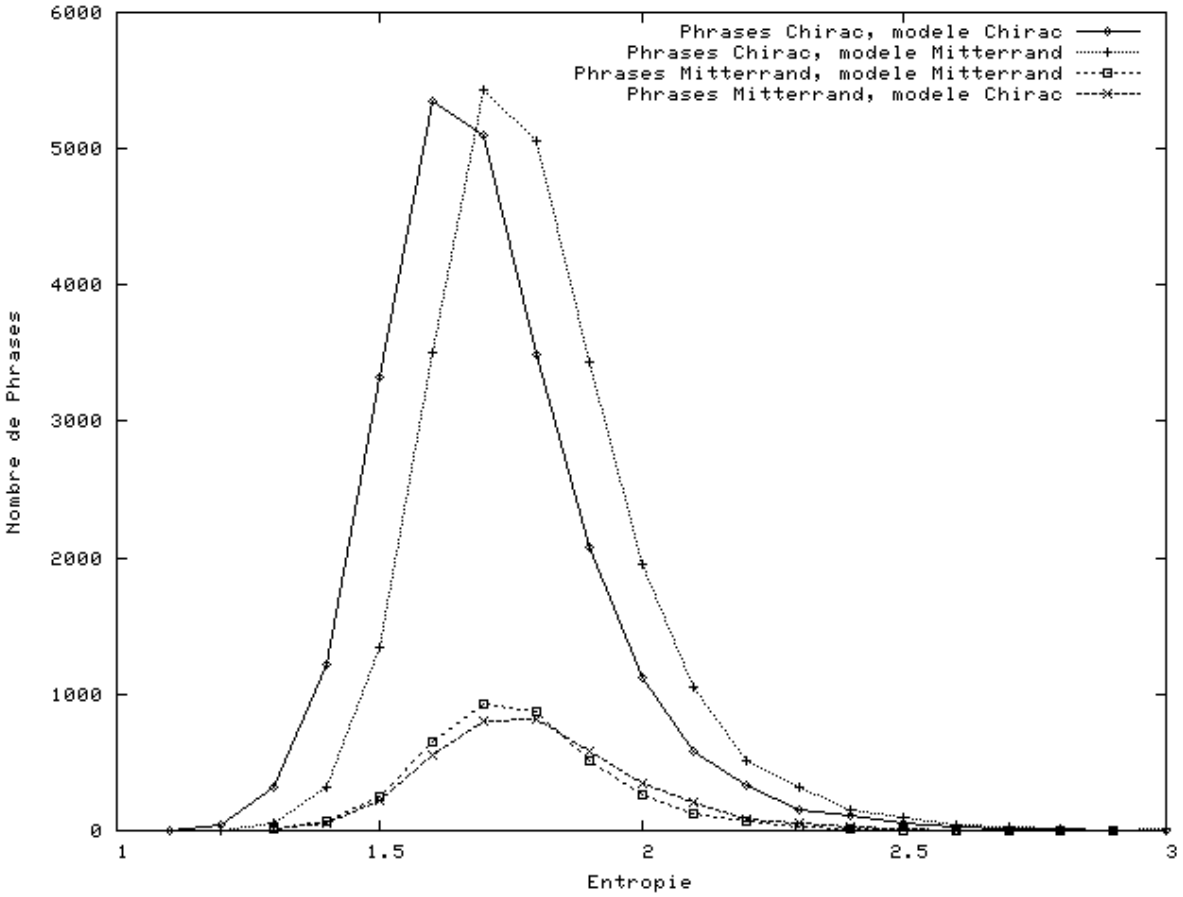

Figure 2 : Nombres de phrases de Chirac et de Mitterrand du corpus de test en fonction des valeurs d'entropie de ces phrases calculées avec les deux modèles 4 -grammes de caractères de Chirac et Mitterrand

Les deux courbes les plus hautes sont associées aux phrases de Chirac, la plus à gauche est donnée par le modèle de Chirac, la valeur moyenne est 1,7. La courbe à droite donnée par le modèle Mitterrand a une valeur moyenne d'entropie plus forte, 1,8, et se distingue nettement de l'autre courbe. Par contre les deux courbes plus basses associées aux phrases de Mitterrand se distinguent très peu avec une entropie moyenne de 1,8. On peut en déduire que les phrases de Chirac sont identifiées plus facilement que les phrases de Mitterrand, ce qui est confirmé par la F-mesure de reconnaissance des phrases de Chirac qui est 0,97. La différence de taille entre les données d'apprentissage pour Chirac et Mitterrand pénalise notablement le modèle Mitterrand.

4. Conclusion

Nous avons montré qu'un modèle de langage probabiliste fondé sur des n-grammes de caractères permet de repérer dans le discours d'un auteur des phrases d'un autre auteur avec une efficacité comparable à celles de modèles plus complexes et mettant en oeuvre d'autres traits caractéristiques statistiquement moins robustes. Cette méthode est simple, efficace et rapide à mettre en œuvre. Les caractères incluant lettres, chiffres et ponctuations semblent être des traits discriminants des auteurs au niveau de la phrase. La disproportion entre les

données d'apprentissage pour les deux auteurs est un handicap important pour une bonne identification.

Références

Alphonse E., Amrani A., Azé J., Heitz T., Mezaour, A.D. et Roche M. (2005). Préparation des données et analyse des résultats de DEFT'05. *Actes TALN & RECITAL*, vol 2 : 99-111.

Clarkson, P.R. and Rosenfeld R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. *Proc of ESCA Eurospeech*, Rhodes, Grèce.

Cover T.M. and Thomas J.A. (1991). *Elements of Information Theory*. J. Wiley Éd, New-York.

Hurault-Plantet M., Jardino M. et Illlouz G. (2005). Modèles de langage n-grammes et segmentation thématique pour une tâche de filtrage de textes. *Actes TALN & RECITAL*, vol 2 : 135-144.

Juola, P. (1998). Cross-Entropy and Linguistic Typology. In D. Powers (ed), *Proceedings of New Methods in Language Processing*, 3, Sydney, Australie.

Khmelev, D.V. and Tweedie, J.T. (2002). Using Markov Chains for Identification of Writers. *In Literary and Linguistic Computing*, vol.16, n°4 : 299-307.

Markov, A.A. (1913). An example of Statistical Study on the Text of Eugene Onegin illustrating the linking of events to a chain (titre traduit du russe). *Izvestija Imp. Akademii nauk*, serija VI, 3 : 153-162.

Peng F., Schuurmans D., Keselj V. et Wang S. (2003). Language Independendent Authorship Attribution using Character Level Language Models. *Proc. of ACL 2003*, Sapporo, Japon.

Pierron L., Coskun D. et Chevalier J-B. (2005). Classification, combinaison et regroupements pour séparer les discours de Mitterrand de ceux de Chirac. *Actes TALN & RECITAL*, vol 2 : 165-173.

Shannon C.E. (1951). Prediction and entropy of printed English. *Bell Syst.Techn* : 50-64.

Teahan, W.J. (2000). Text classification and segmentation using minimum cross-entropy. Proc. of RIAO 2000, Paris, France.

Witten I.H. and T.C. Bell (1991). *The zero-frequency problem : estimating the probabilities of novel events in adaptative text compression*. IEEE Transactions on Information Theory : 37-4.