Modèles de langage n-grammes et segmentation thématique pour une tâche de filtrage de textes

Martine Hurault-Plantet (1), Michèle Jardino (1), Gabriel Illouz (1)

LIMSI-CNRS
Bât.508, Université Paris XI, 91403, Orsay
{Martine.Hurault-Plantet, Michèle.Jardino, Gabriel.Illouz}@limsi.fr

Mots-clés : filtrage de textes, modèle de langage n-grammes, segmentation thématique
Keywords: text filtering, n-gram language model, topic segmentation

Résumé La tâche soumise à évaluation dans l'atelier DEFT'05 consistait à identifier les phrases issues d'allocutions de François Mitterrand, qui avaient été préalablement insérées dans un ensemble d'allocutions de Jacques Chirac. Dans chaque document, le thème des phrases insérées et le thème de l'allocution de Chirac dans laquelle elles s'insèrent, sont différents. Dans cet article, nous présentons les méthodes utilisées au LIMSI pour résoudre cette tâche. Nous avons expérimenté deux méthodes automatiques différentes que nous avons ensuite fait coopérer. L'une de ces deux méthodes s'appuie sur des modèles de langage n-grammes, l'autre est basée sur la segmentation thématique de l'allocution.

Abstract The task for the DEFT'05 evaluation workshop consists in identifying sentences selected from speeches by François Mitterrand, which had been inserted within a set of speeches by Jacques Chirac. The topic of the inserted sentences and the topic of the Chirac speech differ one from another in each document. This paper presents the two methods experimented at LIMSI in order to solve this task. The first one is based on n-gram language models and the second one is based on topic segmentation. Results of both methods are then merged.

1 Introduction

La tâche soumise à évaluation dans l'atelier DEFT'05 consistait à identifier les phrases issues d'allocutions de François Mitterrand, qui avaient été préalablement insérées dans un ensemble d'allocutions de Jacques Chirac. Dans chaque document, le thème des phrases insérées et le thème de l'allocution de Chirac dans laquelle elles s'insèrent, sont différents. Dans cet article, nous présentons les différentes méthodes utilisées au LIMSI pour résoudre cette tâche. Nous avons expérimenté deux méthodes automatiques différentes que nous avons ensuite fait coopérer. L'une de ces deux méthodes s'appuie sur des modèles de langage n-grammes (Jelinek, 1998), l'autre est basée sur la segmentation thématique de l'allocution. Le système

de coopération entre les deux méthodes utilise principalement l'intersection entre les résultats obtenus par chacune des deux méthodes.

Dans la première méthode, nous avons construit des modèles de langage n-grammes appris à partir des corpus d'apprentissage fournis, un pour Chirac et un pour Mitterrand. Chacun des modèles a ensuite été appliqué sur chacune des phrases à tester, l'auteur reconnu est celui dont le modèle donne la plus grande probabilité à la phrase.

La deuxième méthode est basée sur la segmentation thématique : le système détermine, pour chaque allocution, un ensemble de thèmes à partir des mots les plus fréquents. Pour chaque thème, le système calcule les intervalles de phrases de l'allocution dans lesquels on le trouve. Deux pôles thématiques sont ensuite sélectionnés parmi les thèmes : le thème le plus fréquent et son complémentaire, sur le critère de non recouvrement des intervalles associés à chaque thème. Chacun des autres thèmes est ensuite agrégé au pôle avec lequel il a un intervalle commun. Les thèmes qui ont un intervalle commun avec les deux pôles à la fois ne sont pas agrégés. Le système prend pour thème de Chirac celui des deux pôles dont les intervalles sont les plus proches du début et de la fin de l'allocution, et pour thème de Mitterrand le pôle complémentaire.

Dans la suite de l'article, nous présentons successivement les deux méthodes puis la méthode de fusion des résultats. Nous conclurons sur un bilan de ces méthodes et les expérimentations que nous envisageons pour améliorer nos résultats.

2 Les modèles de langage n-grammes

Ces modèles sont généralement employés dans les systèmes de reconnaissance de la parole (Jelinek, 1998). Ce sont des modèles probabilistes qui prédisent un mot connaissant les n-1 mots précédents. Nous les avons utilisés après avoir testé des représentations du type « sac de mots » qui ont donné des résultats peu satisfaisants.

2.1 Prétraitement des phrases et approche « sac de mots »

Nous avons effectué un premier prétraitement simple : nous avons conservé la ponctuation qui est un bon indicateur du style et transformé toutes les majuscules en minuscules.

Ensuite nous avons utilisé un algorithme de classification non supervisée (Jardino, 2000) pour partager toutes les phrases de la tâche 1 en deux sous-ensembles. L'espace de représentation est celui des mots indépendamment de leur ordre. Cette méthode appliquée à un corpus bien constitué (Brown Corpus) avait donné d'excellents résultats (Illouz et al., 2001). Sur le corpus Chirac-Mitterrand les résultats sont nettement moins bons. Le tableau 1 reporte les valeurs de précision et rappel des phrases de Mitterrand obtenues sur le corpus d'entraînement initial de la tâche 1, en utilisant différents ensembles de mots. Dans tous les cas la précision est très mauvaise. Les choix des fréquences permettent d'expérimenter différentes zones de la courbe de Zipf comme les zones de haute fréquence où se regroupent les mots-outils ou les zones de moyenne fréquence, censées représenter les mots sémantiquement significatifs. De manière caricaturale nous avons également expérimenté une représentation réduite au point et à la virgule.

Ensemble de mots	Rappel (%)	Précision (%)
Tous les mots	66	19
Mots de fréquence > 500	50	18
Point et Virgule	50	12
Mots de fréquence < 500	59	17
Mots tels que 10 < fréquence < 500	57	16

Tableau 1 : Rappel et précision des phrases de Mitterrand pour une partition non supervisée en deux classes des phrases de la tâche1

Les meilleurs résultats sont ceux obtenus avec tous les mots. En conséquence, nous les avons conservés et utilisés dans un modèle plus puissant prenant en compte l'ordre des mots dans la phrase.

2.2 Modèles de langage n-grammes pour identifier l'auteur de chaque phrase

Nous avons utilisé le logiciel de CMU (Clarkson, 1997) pour construire les modèles de langage n-grammes P_C et P_M à partir des comptes des successions de n mots respectivement dans les phrases de Chirac et dans les phrases de Mitterrand. Si une phrase de longueur l est représentée par la succession des mots : $m_1...m_i...m_l$, la probabilité de cette phrase calculée à partir d'un modèle n-grammes pour l'auteur A est :

$$P_A(phrase) = \prod_{i=1}^{i=l} p_A(m_i / m_{i-n+1}...m_{i-1})$$

Nous avons utilisé un lissage de type Witten-Bell pour calculer les probabilités des évènements non observés dans le corpus d'apprentissage. Ce type de lissage prend en compte le nombre de contextes dans lesquels ont été observés les n-grammes dans le corpus d'apprentissage. Pour choisir la valeur de n la mieux adaptée aux données, nous avons partitionné le corpus de la tâche1 de la façon suivante : 90% pour le corpus d'apprentissage des modèles Chirac et Mitterrand et 10% pour le corpus de test. Nous avons calculé pour chaque phrase du test les probabilités $P_C(phrase)$ et $P_M(phrase)$ et attribuée la phrase à l'auteur dont le modèle donne la plus grand probabilité. Nous avons ensuite évalué les valeurs de précision et de rappel pour les phrases de Mitterrand en comparant hypothèses et références, ceci pour des valeurs de n allant de 2 à 4.

Le tableau 2 montre que le modèle 3-grammes donne les meilleurs résultats sur nos jeux de données.

Modèles	Rappel (%)	Précision (%)
2-grammes	61	40
3-grammes	58	78
4-grammes	58	44

Tableau 2 : Rappel et précision des phrases de Mitterrand dans le corpus de test pour différents modèles de langage

2.3 Améliorations possibles

Une seule partition du corpus en apprentissage (90%) et test (10%) a été effectuée par manque de temps. D'autres tests ont été réalisés après l'expérimentation initiale, montrant des performances plus faibles en terme de précision pour les modèles 3-grammes. Les très bons résultats obtenus par ces derniers sont à relativiser par la sélection d'un corpus de test qui n'était pas en fait assez représentatif.

Les modèles de prédiction peuvent être améliorer par des méthodes de ré-échantillonnage en faisant varier phrases de test et d'apprentissage par validation croisée ou par la technique de Jackknife (Lebart et al., 2000).

2.4 Lissage pour construire des ensembles continus de phrases d'un même auteur

Notre méthode vote pour chaque phrase, ce qui induit des passages discontinus de phrases de François Mitterrand dans les phrases de Jacques Chirac. Pour pallier cet effet, nous avons utilisé un algorithme simple de lissage : chaque fois qu'un ensemble de 1 à k phrases de Jacques Chirac est détecté entre deux phrases de Mitterrand, ces phrases sont attribuées à François Mitterrand. Cette méthode donne un lissage assez fruste avec un taux de rappel très important. La valeur k = 4 a donné les meilleurs résultats sur le corpus de test.

3 Les pôles thématiques de l'allocution

La deuxième méthode présentée est basée sur une segmentation thématique de chaque allocution. Plus précisément, le système détermine d'abord l'ensemble des thèmes de l'allocution, puis, parmi ces thèmes, le thème dominant et le thème qui s'en éloigne le plus. Les méthodes décrites dans cette section utilisent un certain nombre de seuils qui ont été déterminés empiriquement sur le corpus d'apprentissage.

3.1 Détermination des thèmes

Pour déterminer les thèmes d'une allocution, nous effectuons d'abord un pré-traitement des phrases. Les mots de chaque phrase sont lemmatisés (Schmid, 1999), et nous ne conservons

pour indexer la phrase que les lemmes (ou le mot lorsqu'il est inconnu) des substantifs, des adjectifs, des noms propres et des abréviations. Nous utilisons également une liste classique de mots vides[1] à laquelle nous avons ajouté les treize mots les plus fréquents du vocabulaire du corpus d'apprentissage ainsi que les mots *monsieur* et *président*. Nous effectuons ensuite une reconnaissance automatique des adjectifs dont le substantif est dans l'allocution, et nous remplaçons ces adjectifs par les substantifs correspondants. Pour effectuer la reconnaissance automatique des adjectifs, le système utilise une liste de terminaisons d'adjectifs[2] qui lui permet de déterminer des adjectifs candidats et leurs racines respectives. Le système recherche ensuite les substantifs qui commencent par ces racines. Le bruit généré par cette méthode est limité par deux contraintes : la racine doit avoir une taille minimum (au moins deux caractères), et la taille de l'adjectif doit être supérieure à celle du substantif.

Pour renforcer les thèmes, le système utilise une méthode supplémentaire basée sur les séquences fréquentes maximales (Grahne et al., 2003). Le système recherche d'abord les séquences les plus fréquentes de mots dans une allocution, chaque phrase étant considérée comme une transaction. Le système génère ensuite la règle suivante d'équivalence entre les mots des séquences trouvées : le mot le plus fréquent de chaque ensemble est substituable à chacun des autres mots de la séquence. Le système ré-indexe alors chaque phrase de l'allocution suivant ces règles. Le thème le plus fréquent est donc renforcé par les thèmes qui sont en forte cooccurrence avec lui. Pour limiter le bruit produit, nous avons choisi un support élevé (la cinquième plus forte fréquence) pour générer les séquences fréquentes maximales.

Les thèmes finalement retenus par le système sont les mots les plus fréquents de l'allocution. Nous avons choisi un seuil égal à la vingt-cinquième plus forte fréquence, ou à défaut, un seuil de fréquence égal à 2. Par exemple, pour l'allocution 242 du corpus de test, nous obtenons les phrases indexées suivantes :

C 242:1 présidente flamme cas
C 242:2 émotion vie
C 242:3 présidente mental_handicap madame mental handicap accueil
C 242:4 combat obstacle digne
C 242:5 flamme droit
C 242:6 droit handicap
C 242:7 dignité droit
M 242:8 sujet général mondial attention négociation commerce débat raison
M 242:9 liberté général commerce échange
M 242:10 sujet rencontre mot
M 242:11 général obstacle accord
M 242:12 tiers accord
M 242:13 justice mot traitement
M 242:14
M 242:15 mondial clause accord
M 242:16 clause droit nation accord traité
M 242:17 cas messieurs mesdames

---

[1] Cette liste a été trouvée sur le Web

[2] Idem

M 242:18 traité
M 242:19 nation
M 242:20
M 242:21 justice négociation chemin force
M 242:22 débat
M 242:23 général accord
M 242:24 besoin
M 242:25 mondial
M 242:26
M 242:27 traitement
M 242:28 droit
M 242:29 vrai part
M 242:30 échange
C 242:31 dignité attention droit handicap
C 242:32 regard
C 242:33 handicap
…….

Nous avons trouvé une séquence maximale, *mental_handicap,* ainsi que deux correspondances adjectif-substantif, *handicapé-handicap* et *national-nation*. Le système associe ensuite à chaque thème les intervalles de phrases de l'allocution où il apparaît. Pour lisser les intervalles, le système calcule la densité moyenne du thème sur des groupes de cinq phrases, à partir de la première phrase[3]. Ainsi, si un même thème apparaît deux fois à une distance de trois phrases, le système considère qu'il apparaît dans un intervalle de cinq phrases qui comprend les deux phrases où il apparaît et les trois phrases où il n'apparaît pas. Les limites exactes de chaque intervalle sont ensuite déterminées par la recherche de la première et de la dernière phrase de l'intervalle où le thème apparaît.

Les thèmes les plus fréquents de l'allocution 242 et leurs intervalles sont les suivants (les phrases sont re-numérotées à partir de 0) :

12	handicap	2_5 30_32 41_68
7	dignité	6_6 30_37 47_61
6	droit	4_6 15_15 27_30
6	vie	1_1 36_36 48_64
5	accord	10_22
4	combat3_3 37_39	

3.2 Détermination des pôles thématiques

Le pôle dominant est le thème le plus fréquent. Le système cherche ensuite le pôle complémentaire. Deux méthodes ont été expérimentées pour cette recherche. Dans la première méthode, le système recherche le premier thème dont les intervalles n'ont aucun recouvrement avec les intervalles du pôle dominant. Si le système n'a pas trouvé de thème complémentaire, il applique une deuxième méthode qui consiste à rechercher le thème qui a le moins d'intervalles en commun avec le thème dominant. Pour cela, le système ré-indexe

[3] Pour les allocutions courtes (moins de 35 phrases), nous utilisons des groupes de trois phrases.

d'abord chaque thème par les thèmes qui possède un intervalle commun avec lui. Puis il recherche le thème qui a à la fois le moins de thèmes en commun et le plus de thèmes différents de ceux qui indexent le thème dominant.

Dans l'exemple de l'allocution 242, le pôle dominant est *handicap,* le pôle complémentaire est *accord*, premier thème dont l'intervalle n'a aucune intersection avec le pôle dominant.

3.3 Détermination des phrases de Mitterrand

Une fois les deux pôles thématiques trouvés, le système agrège chacun des autres thèmes avec le pôle avec qui il possède un intervalle en commun. Si un thème a un intervalle en commun avec les deux pôles, il est considéré comme un thème commun aux deux pôles et n'est pas agrégé.

Dans l'exemple de l'allocution 242, le pôle *handicap* est agrégé avec les thèmes *dignité*, *vie*, *combat*, *vrai*, *regard*, *besoin*. Le pôle *accord* est agrégé avec les thèmes *général*, *mondial*, *négociation*, *nation*, *justice*, *échange*. Les thèmes *droit*, *raison*, *obstacle*, *rencontre*, *liberté* sont des thèmes communs aux deux pôles, ils ne sont donc pas agrégés. Le système produit finalement les intervalles suivants :

Pôle *handicap* : intervalles 0_6 23_23 28_28 30_39 41_69
Pôle *accord* : intervalles 7_22 24_24 26_26 29_29

Le système attribue le locuteur Chirac à celui des deux pôles qui débute l'allocution et la termine. Il attribue le locuteur Mitterrand à l'autre pôle. Le système indexe ensuite les phrases des intervalles associés à chacun des deux pôles par leurs interlocuteurs respectifs, à l'exception des intervalles ne comportant qu'une seule phrase. Nous estimons en effet que, si le thème est isolé dans une seule phrase, il peut s'agir d'un thème en partie commun aux deux locuteurs. Par ailleurs, si les intervalles finaux des pôles (pôle et thèmes agrégés) se croisent, c'est-à-dire si l'un débute l'allocution et l'autre la termine, nous considérons que les deux pôles sont des thèmes de Chirac, l'un étant un sous-thème de l'autre, et le système indexe toutes les phrases par le locuteur Chirac.

Les phrases indexées de l'allocution 242 sont donc dans les intervalles suivants :

Phrases attribuées à Chirac : intervalles 0_6 30_39 41_69
Phrases attribuées à Mitterrand : intervalles 7_22

Certaines phrases ne sont pas indexées, soit parce qu'elles ne contiennent aucun des thèmes retenus, soit parce qu'elles ne contiennent qu'un thème commun aux deux pôles. Finalement, le système retient comme phrases de Mitterrand les phrases des intervalles du pôle que le système a étiqueté Mitterrand ainsi que les phrases non indexées qui sont comprises entre les limites d'un intervalle Mitterrand et la limite de l'intervalle Chirac qui suit et de celui qui précède. Ce qui donne l'intervalle 7_29 (c'est-à-dire les phrases <242 :8> à <242 :30>) pour les phrases de Mitterrand dans l'allocution 242.

3.4 Les limites de la méthode

La méthode de détermination des pôles thématiques ne s'applique pas toujours aussi bien que dans l'exemple de l'allocution 242, et cela pour plusieurs raisons. Tout d'abord, la reconnaissance des thèmes repose sur les mots les plus fréquents de l'allocution. Or il se trouve que les deux locuteurs ont souvent, à l'intérieur d'une même allocution, un grand nombre de mots en commun. Il arrive même que les phrases de Mitterrand n'aient aucun mot spécifique fréquent, ou que le thème dominant soit un thème commun entre Chirac et Mitterrand. Par ailleurs, si on arrive en général bien à déterminer le thème dominant, en grande majorité attribué à Chirac, en revanche nous n'avons pas trouvé de méthode sûre nous permettant de distinguer entre le thème de Mitterrand et un sous-thème de Chirac. C'est plus particulièrement le cas lorsque l'allocution de Chirac est longue. L'exemple de l'allocution 111 du corpus de test illustre bien ces problèmes.

Pôle dominant *europe* : intervalles 24_44 78_84 99_104 110_119 127_127
Pôle complémentaire *développement* : intervalles 50_50 69_69 131_139

Les deux thèmes principaux de Mitterrand dans cette allocution sont en réalité *exploitation* (intervalle 94_100) et *agricole* (intervalle 94_105), et le thème *développement* est en réalité un sous-thème de l'allocution de Chirac. Le problème vient de ce que le pôle dominant trouvé est un thème commun aux allocutions de Chirac et de Mitterrand : *europe* apparaît en effet à la fois dans l'allocution de Chirac et dans l'intervalle 99_104 des phrases de Mitterrand.

Parmi les 294 allocutions du corpus de test, on a un grand nombre de pôles qui sont des thèmes communs aux deux locuteurs. Le tableau 3 récapitule les différentes combinaisons de locuteurs réels trouvées dans nos résultats.

Locuteur réel du pôle dominant	Locuteur réel du pôle complémentaire	Nombre d'allocutions
Chirac	Mitterrand	67
Mitterrand	Chirac	9
Chirac	Mitterrand et Chirac	45
Mitterrand *et* Chirac (*ou* Chirac *ou* Mitterrand)	(Chirac *ou* Mitterrand *ou*) Mitterrand *et* Chirac	107
Chirac	Pas de pôle complémentaire trouvé	42
Chirac	Chirac	69

Tableau 3 : Les locuteurs réels des pôles thématiques trouvés

4 La coopération entre les deux méthodes

La méthode des modèles de langage n-grammes et la méthode de segmentation thématique produisent chacune un ensemble de phrases attribuées à Mitterrand. Les deux méthodes obtiennent des Fscores proches mais avec des précisions et rappels différents. La méthode utilisant les modèles de langage obtient un très bon rappel mais une précision médiocre. La méthode utilisant la segmentation thématique obtient une meilleure précision mais un plus mauvais rappel. La stratégie de fusion des résultats adoptée consiste à retenir d'une part les phrases présentes dans les deux résultats afin d'augmenter la précision, et d'autre part les phrases de la méthode utilisant les modèles de langage lorsqu'on obtient aucune phrase par l'autre méthode, afin de garder un bon rappel. Nous avons obtenu une légère amélioration des performances. Par exemple, la fusion des résultats pour la tâche 3 donne les résultats suivants :

Modèles de langage : précision = 0.41, rappel = 0.88, Fscore(beta=1) = 0.56
Segmentation thématique : précision = 0.52, rappel = 0.55, Fscore(beta=1) = 0.53
Fusion des deux méthodes : précision = 0.51, rappel=0.66, Fscore(beta=1) = 0.57

5 Conclusion

Au vu des résultats obtenus, nos méthodes ont incontestablement des faiblesses. L'apprentissage du modèle de langage de chaque auteur doit être amélioré par des méthodes de ré-échantillonnage (voir paragraphe 2.3). Par ailleurs, l'existence de nombreux thèmes communs entre les auteurs rend faiblement efficace la segmentation de l'allocution en deux pôles thématiques (voir paragraphe 3.4). Pour résoudre l'ambiguïté créée par les thèmes communs aux deux auteurs, d'autres exploitations des thèmes sont envisagées. En particulier, nous n'avons pas utilisé le positionnement des thèmes les uns par rapport aux autres dans le fil du discours. En effet, certains thèmes sont entrelacés le long du discours alors que d'autres se séparent plus nettement. La fusion des résultats des deux méthodes produit une amélioration assez faible (voir paragraphe 4). Nous envisageons d'étudier le recoupement entre les thèmes trouvés et les prédictions des modèles de langage pour chaque allocution, afin de trouver un autre type de coopération entre les deux méthodes.

Références

CLARKSON P.R., ROSENFELD R. (1997), Statistical Language Modeling Using the CMU-Cambridge Toolkit, Actes de *ESCA Eurospeech 1997*.

GRAHNE G., ZHU J. (2003), Efficiently Using Prefix-trees in Mining Frequent Itemsets, Actes de *First IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Melbourne, FL.

ILLOUZ G., JARDINO M. (2001), Analyse statistique et géométrique de corpus textuels, *T.A.L., Traitement automatique des langues et linguistique de corpus,* Vol 42:2, pp. 501-516.

JARDINO M. (2000), Unsupervised non-hierarchical entropy-based clustering, *Data Analysis, Classification and Related Methods,* Eds. H.-H.Bock, W.Gaul, M.Schader. Springer.

JELINEK F. (1998), *Statistical Methods for Speech Recognition* , MIT Press.

LEBART L., MORINEAU A., PIRON M. (2000), *Statistique exploratoire multidimensionnelle*, Dunod.

SCHMID H. (1999), Improvements in Part-of-Speech Tagging with an Application To German. In Armstrong, S., Chuch, K. W., Isabelle, P., Tzoukermann, E. & Yarowski, D. (Eds.), *Natural Language Processing Using Very Large Corpora*, Dordrecht, Kluwer Academic Publisher.