Un modèle HMM pour la détection des mots composés dans un corpus textuel.

Lakhdar Remaki & Jean Guy Meunier

LANCI
Université du Québec à Montréal.

Cp 8888, succ A
Montréal. H3C 3P8
Canada
remaki@livia.etsmtl.ca meunier.jean-guy@uqam.ca

Abstract

We propose in this paper a HMM-based approach for complex word detection in French language. This system is composed of two complementary sub-models, one related to the corpus, and the other is related to the field. A very encouraging preliminary results are then presented.

Résumé

Nous proposons dans cet article un modèle basé sur les chaînes de Markov cachées (HMM) pour l'analyse des mots composés d'un texte. Le système construit est constitué de deux sous modèles complémentaires, un relié au texte et l'autre au domaine. Les premiers résultats, que nous présentons dans cet article, sont très satisfaisants et encourageant pour une amélioration du système.

Mots-clés : N-Grams, HMM, Mots composés, collocation, cooccurrence.

1. Introduction

- **Le dépouillement des mots composés.**

Il existe de nombreux systèmes de dépouillement terminologique dont une des tâches est d'identifier les syntagmes complexes (synopsies ,mots composés, collocation etc.) La grande majorité de ces systèmes opèrent a partir d'analyses linguistiques . Bourigault, 1992, Jacquemin 1994). Mais dans la littérature technique on voit de plus en plus apparaître des systèmes dont le traitement est de nature plus statistique (Daille. 1994, Chruch et all. 1990 etc,). Dans cette approche, les systèmes produisent des ensembles ou listes de collocations, c' est-à-dire des séquences d'unité d'information (mots ou n-grams) sur lesquelles certains critères de sélection sont appliqués, tels par exemple , Chi carré, seuil bayésien, ratio probabiliste, information mutuelle etc. Malgré la simplicité de ces méthodes, ces approches statistiques semblent donner des résultats relativement intéressants. (Manning et all 1999).

Ce type d'analyse sert plusieurs fins. Outre l'assistance au dépouillement terminologique, l'identification de mots composés améliore de manière significative, la génération automatique de langage naturel, (Smajda 1993) le rappel d'information, la catégorisation automatique des textes, la comparaison de textes multilingues, etc.

L'approche que nous explorons dans cette recherche est dans ce même esprit. Elle est de nature statistique mais en appelle a la théorie des chaîne cachées de Markov (Hidden Markov Models HMM). Cette approche permet un traitement similaire sur toutes les langues , mais surtout , et la est un point important, elle permet un apprentissage des mots composés significatifs pour un domaine particulier et un utilisateur spécifique. En effet, contrairement aux autres approches qui produisent une liste générale de mots composés sur un corpus donné, l'approche HMM met en jeu un mémoire de séquences acceptées et pertinentes pour un utilisateur . Ceci permet alors à chaque utilisateur et selon la nature des corpus traités de produire des dictionnaires très réduits appuyés par la connaissance des mesures de probabilités appropriées lui permettant d'accéder automatiquement à la plus des mots composés propre a son domaine de spécialité.

Dans la présente recherche , nous explorons la valeurs de cette approche sur le plan conceptuel , en présentons une application et une expérimentation sur un corpus.

- **Le modèle HMM**

Un modèle HMM (Hidden Markov Model) est un modèle stochastique particulier, proposé pour la première fois à la fin des années 60 début des années 70 (Baum. 1970, Baum. 1972, Baker. 1975a, Baker. 1975b). Il représente un objet donné par deux suites de variables aléatoires : l'une dite caché et l'autre observable. La suite caché correspond à la suite d'états q_1, q_2, ..., q_T, noté Q(1 : T), où les q_i puisent leur valeur parmi l'ensemble des N états du modèle { s_1, s_2,..., s_N}. La suite observable correspond à la suite d'observations o_1, o_2,..., o_T, notée O(1 : T), où les o_i sont aussi fonctions du temps et se réalisent parmi un ensemble M de symboles observables { v_1, v_2,..., v_M}. Les modèles HMM, sont très utilisés dans le domaine de la reconnaissance, particulièrement en parole et imagerie. La philosophie de la méthode consiste à identifier un objet donné moyennant un ensemble d'observations, en d'autres termes : soit O la suite d'observations réalisée lors de la présentation d'une forme ou d'un objet à identifier, au modèle, on cherche à maximiser la probabilité p(Q/O), ou Q et une suite d'états. La connaissance de la matrice des probabilités de transitions A, et la matrice d'observation des symboles B, permet d'estimer cette probabilité par la formule de Bayes. La maximisation se fait donc sur tout les chemins Q, possibles. Ainsi le chemin optimal nous permettra d'identifier notre objet, d'ou l'importance du choix d'états dont toute suite permise par le modèle soit facilement interprétable.

Nous proposons dans ce travail deux modèles HMM. L′un lié au corpus lui même, que nous désignerons par HMM_{corpus}, et l′autre lié au domaine que nous désignerons par $HMM_{domaine}$. Le premier est responsable de détecter les mots composés du corpus en utilisant celui ci comme seule source d′information, et le second sera lié au domaine et apprendra à partir de celui ci et de façon continue. Sont rôle est donc d′être un outil complémentaire au premier et apportera sa concurrence pour améliorer les résultats. Bien sur cette apport dépendra de l′état d′apprentissage de ce HMM, la Figure 3 schématise la complémentarité de ces deux outils .

2. Description du modèle

Passons maintenant à description du modèle, il s'agit donc de terminer l'ensemble des états S, et l'ensemble des observations V. Dans notre approche et pour le cas HMM_{corpus} l'ensemble S

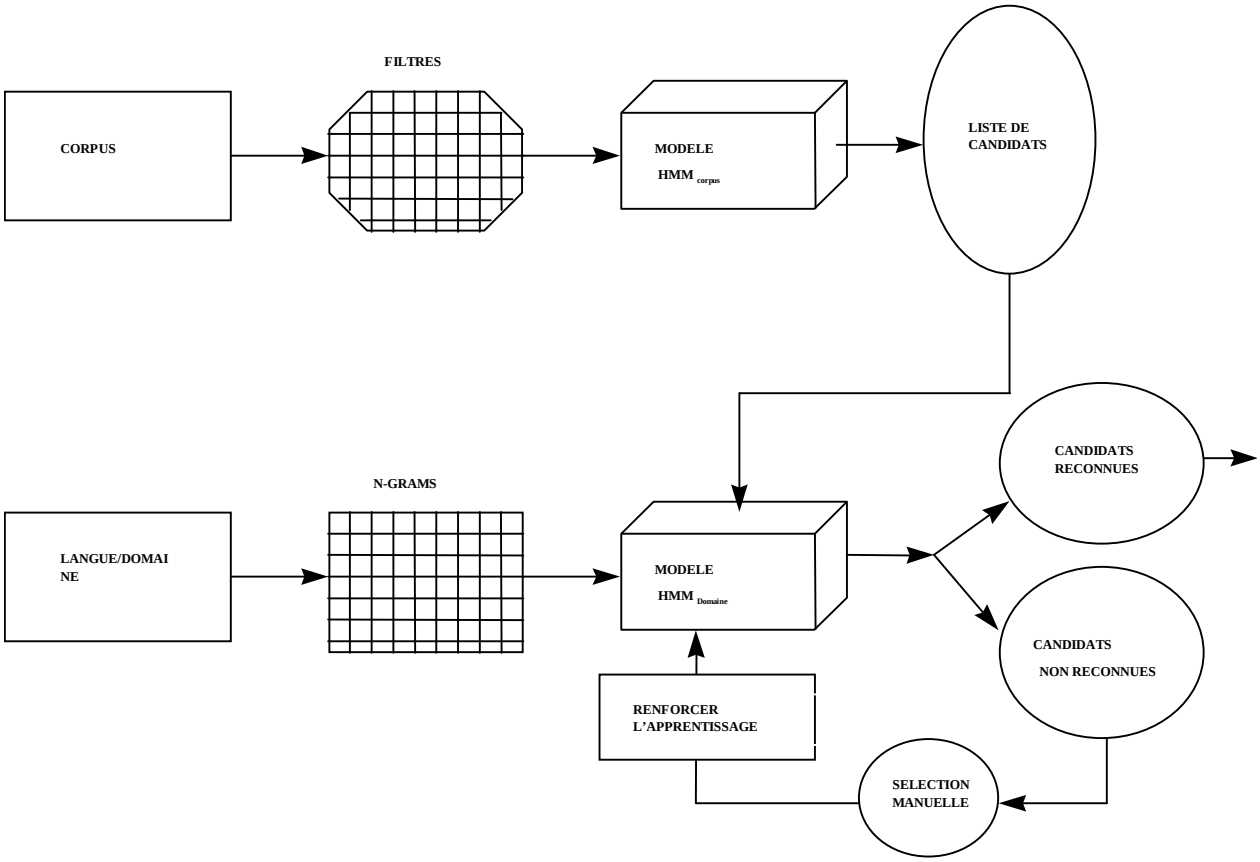

Figure 3 : Schéma représentatif du système HMM_{corpus}- $HMM_{domaine}$

3. Expérimentation

Le tableau ci dessous représente les premiers résultats de l'application du modèle HMM_{corpus} décrit plus haut. Nous avons appliqué le texte sur les corpus suivants .

Zola….
Papon….
Renault …..

Nous sommes actuellement à l'appliquer à un texte de 1000 pages. .

Tableau 1 : Résultats préliminaires du détecteur des mots composés

Texte	Nombre de pages	Recall	Precision
Zola.txt	7	50%	64%
Papon.txt	7	65%	55%
Renault.txt	40	92.4%	75%

4. Discussion

Bien que notre corpus soit encore limité nous trouvons des résultats intéressants. De fait, la performance de ce modèle croît avec la taille du corpus et la spécialisation de celui-ci dans un domaine donné, et la les taux de réussite, comme le montre le tableau atteignent des proportions exceptionnelles. Ici uniquement le HMM_{corpus} est utilisé, la contribution du $HMM_{domaine}$ améliorera encore d'avantage ces résultats, (lors de réalisation de ce testes l'apprentissage du $HMM_{domaine}$ n'était pas encore complet chose qui nous nous a pas permise de faire des tests avec la compétition de celui-ci).

Références

Baker J.K. (1975a), Stochastic modeling as a means of automatic speech recognition. Ph.D. Dissertation, Carnegie-Mellon Univ. 1975.

Baker J.K. (1975b), The Dragon system- an overview. IEE Trans. Acoustics, Speech and Signal Processing. ASSP-23, pp.24-29, 1975.

L. E. Baum, Ted Petrie, George Soules, and Norman Weiss. (1970), Amaximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The Annals of Mathematical Statistics,41(1):164-171, 1970.

L. E. Baum. (1972), An inequality and associated maximization technique in statistical estimation for probabilistic functions of a markov process. Inequalities, 3:1-8,1972.

Bourigault, D. (1992), Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING'92: Proceedings of the 15th International Conferences on Computational Linguistics*, August 1992, Nantes, France.

Church, K. Hanks. (1990), Word Association norms, mutual information, and lexicography. Computational Linguistics. vol 16 no 1 22-29.

Daille B. Gaussier.E; Langé, JM . (1994), Towards Automatic extraction of Monolingual and Bilingual Terminology, " COLING 94 Kyoto Japan.

Jacquemin, C., ASTR. L, Baker J.K. (1994), An unification based front end to automatic indexing, " Proc. Intelligent Multimedia Information Retrieval System and Management RIAO., New York 1994

Lebart L., salem. A Baker J.K. (1994), Statistique textuelle . Dunod. Paris 1994.

Smajda F. (1992), Retrieving collocations for texts. Z tract6s: Computaional Linguistics. 19: 177. 1992

Manning, C. & Schûtze. C. (1999), Foundations of Statiscial Nautral Language Processing.MIT Press.

B. Teufel. (1989), Informationsspuren zum numerischen und graphischen Vergleich von reduzierten natürlichsprachlichen Texten, Informatik-Dissertation Vdf-Verlag, Zürich Nr. 13, 1989.