Memory-based-Learning et Base de règles pour un Etiqueteur du Texte Arabe

Yamina TLILI-GUIASSA
Laboratoire de Recherche en Informatique
Université Badji Mokhtar, Annaba Algerie
Mel : guiyam@yahoo.fr

Mots clés – Key words :

Etiquetage, Memory-Based Leaning, K-NN, Base de règles, Morphosyntaxique, Langue Arabe.

Tagging, Memory-based learning, K-NN, Based-rules, Morphosyntaxitic, Arabic language.

Résumé – Abstract

Jusqu'a présent il n'y a pas de système automatique complet pour l'étiquetage du texte arabe. Les méthodes qu'elles soient basées sur des règles explicites ou sur des calculs statistiques, ont été développées pour pallier au problème de l'ambiguïté lexicale. Celles-ci introduisent des informations sur le contexte immédiat des mots, mais font l'impasse sur les exceptions qui échappent aux traitements. L'apparition des méthodes Memory-Based Learning(MBL) a permis l'exploitation automatique de la similarité de l'information contenue dans de grandes masses de textes et, en cas d'anomalie, permet de déduire la catégorie la plus probable dans un contexte donné, sans que le linguiste ait à formuler des règles explicites. Ce papier qui présente une approche hybride combine les méthodes à base de règles et MBL afin d'optimiser la performance de l'étiqueteur. Les résultats ainsi obtenus, présentés en section 6, sont satisfaisants et l' objectif recherché est atteint .

Since now there is no complete automatic system for tagging an Arabian text. Methods based on explicit rules or on statistical calculations, have been developed to palliate problems of lexical ambiguousness. They introduce some information on the immediate context of the words but , make the dead end on the exceptions that escape to treatments. The apparition of the Memory-Based Learning(MBL) methods, that exploit automatically the similarity of information contained in big masses of texts and permit, in case of anomaly, to deduct the likeliest category in a given context, without the linguist has to formulate explicit rules. This paper presents an hybrid approach that combines methods based on rules and MBL, thus, in order to optimize the labeller's performance. Our objective is reached and the gotten results, presented in section 6, are satisfactory.

1 Introduction

L'étiquetage morphosyntaxique a pour but d'associer une étiquette grammaticale à chaque mot de la phrase. La première étape est alors de définir un jeu d'étiquettes, adapté au découpage en tronçons, l'étiqueteur morphosyntaxique doit fournir au module de découpage en tronçons toutes les informations grammaticales dont il a besoin pour mener à bien son processus. La majorité des publications sur l'étiquetage automatique font l'impasse sur les exceptions ignorées par les règles morphosyntaxiques, et la plupart des systèmes ont un comportement assez flou, voir inconsistant sur les cas épineux. Particulièrement, la langue arabe présente beaucoup de ces cas, cependant elle est très riche sur le plan morphologique. Certains préfixes, suffixes et des informations fournis par la prise en considération d'un contexte large jouent un rôle capital dans l'analyse morphosyntaxique. Les règles sont correctes pour la majorité des cas, mais la représentation des connaissances de ces règles n'est nécessairement pas parfaite pour tous les cas. Les exceptions ignorées par ces règles doivent être traitées par un autre processus qui prendra en charge les cas épineux avec une grande performance. Pour résoudre ce problème nous proposons une combinaison de la méthode à base de règles et la méthode basée sur l'algorithme des K-plus proches voisins (K-NN)[1]. L'approche proposée garde les K-NN pour chaque erreur commise par la règle. Le but de ce travail est de pallier aux problèmes posés au niveau de découpage en tronçons de la phrase arabe. Ainsi, fondée sur la combinaison de la méthode à base de règle et memory-based learning (MBL)[2], le type d'étiquette est déterminé par la première méthode et sera vérifié par la deuxième. Si le contexte courant est une exception de la règle alors le processus de calcul de similarité est déclenché, cette méthode est efficace pour la prise en charge des exceptions.
L'article présente brièvement l'état de l'art en section 2, traite l'étiquetage morphosyntaxique à base de règles dans la section 3, La section 4 décrit l'étiquetage par MBL et la section 5 explique l'approche proposée. La section 6 expose les résultats obtenus et se termine par une conclusion.

2 Etat de l'art

Le problème rencontré dans les analyseurs syntaxiques traditionnels est celui de la *combinatoire*, qui peut être d'origine lexicale[3] ou structurale[4]. De nouvelles méthodes ont été développées pour pallier à ce problème de combinaison. Fondées sur l'étiquetage morphosyntaxique ou *tagging,* ces méthodes permettent de réduire l'ambiguïté lexicale en introduisant des informations sur le contexte immédiat des mots. Le tagger vient ainsi se substituer à l'analyseur morpho lexical avec comme nouvelle ressource une base de connaissances contextuelles. Ces connaissances peuvent être soit de type probabiliste (si le tagger utilise des informations statistiques sur la contiguïté des mots), soit sous forme de règles explicites.

- bases de règles(TAGGIT de Greene, Rubin, 1971, Francis, Kucera, 1982).
- basées sur les données(Bahl, Mercer,1976, Debili, 1977, Leech et al., 1983, Church, 1988, DeRose, 1988).
- machine learning(TiMBL de Daelemans et al ., 2001).

[1] Les K plus proches voisins
[2] Memory-based learning
[3] plusieurs étiquettes pour un token
[4] plusieurs structures pour une phrase

- méthodes hybrides(Brill, Eric, 1995), Leech et al., 1994, Tapanainen, Voutilainen, 1994, Tzoukermann et al., 1995)

3 L'étiquetage morphosyntaxique à base de règles

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (nom, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît. Malgré l'apparente simplicité de la formulation, l'objectif n'est toujours pas atteint, ou seulement partiellement. La longévité de l'intérêt porté à l'étiquetage témoigne de la difficulté que celui-ci devait en fait receler. Et en même temps de son utilité au regard des applications mettant en œuvre le langage naturel. L'étiquetage de l'arabe hérite de cette situation de fait et voit même sa difficulté s'amplifier lorsque les textes visés se présentent sous leur forme non pas voyelle, mais partiellement seulement, ou encore totalement non voyelle, ce qui correspond au cas le plus courant (Van Mol, 2001).

3.1 Segmentation

Morphologiquement les langues riches, comme l'arabe, présentent des défis significatifs à des applications de traitement de langage naturel parce qu'un mot donne souvent plusieurs significations complexes. L'étiquetage par règles morphosyntaxiques utilise les préfixes et les suffixes comme des identificateurs de catégorie de mot : إستفعلوا: إس# تفعل +وا . La méthode proposée par (L.Young-suk et al., 2002) présente plusieurs avantages pour notre approche ainsi que pour le jeu d'étiquettes adopté(S.Khoja et al., 2001).

3.2 Règles de déduction

1. Les noms
   Le processus pour identifier les noms et les noms propres s'appuie sur les travaux de (S.Abuleil et al., 2002) et (Abuleil, Evens, 1999) respectivement.

2. Les verbes
   La majorité des verbes arabes suivent des règles claires qui peuvent définir leurs morphologies et génère leurs paradigmes, la technique décrite dans (Beesley, Karttunen, 2000) est utilisée dans le système proposé.

3. Les outils
   Ils sont stockés dans une base.

4 L'étiquetage morphosyntaxique par Memory-based learning

Le MBL est décantant direct de l'algorithme K-NN qui utilise des structures de données complexes. Il a un nombre de propriétés intéressantes (W .Daelemans et al., 1996): i) pas de traitement additionnel de lissage pour les données rares. ii) les exceptions peuvent contribuer à une généralisation. (Zavrel, Daelemans, 2000). La similarité entre une instance x et les exemples stockés en mémoire est calculée en utilisant la métrique distance $\Delta(x, y)$:

$\Delta(x, y) = \sum \alpha_i \, \delta(x_i, y_i)$ ou α_i est le poids de $i^{éme}$ attribut , $\delta(x_i, y_i) = 0$ si $x_i = y_i$ et $\delta(x_i, y_i) = 1$ si $x_i \neq y_i$ (Zavrel, Daelemans, 2000, Park, Zhang, 2003)).

5 L'étiquetage morphosyntaxique

L'architecture générale de l'étiqueteur est donnée par la figure 1. Dans la phase apprentissage chaque mot est analysé par les règles, une étiquette est déterminée. L'étiquette déterminée doit être comparer à l'étiquette en entrée, en cas de non égalité alors le mot avec ces deux étiquettes sont stockées dans une liste appelée liste d'anomalies. Durant la classification l'étiquette du mot M_i est déterminée en regardant le mot et le contexte approprié C_i, le type de l'étiquette est ainsi déterminé par le calcul des similarités entre les instances stockés en mémoire et le mot cible (l'étiquette déterminer par les règles est alors écarté). Les informations utilisées pour calculer $\Delta(x, y)$ sont des valeurs qui représentent les mots et les catégories dans une fenêtre à 3 éléments (Hacioglu, Ward, 2003). Les plus importantes sont le mot en question, le tag du mot précédant (M. Diab et al., 2004).

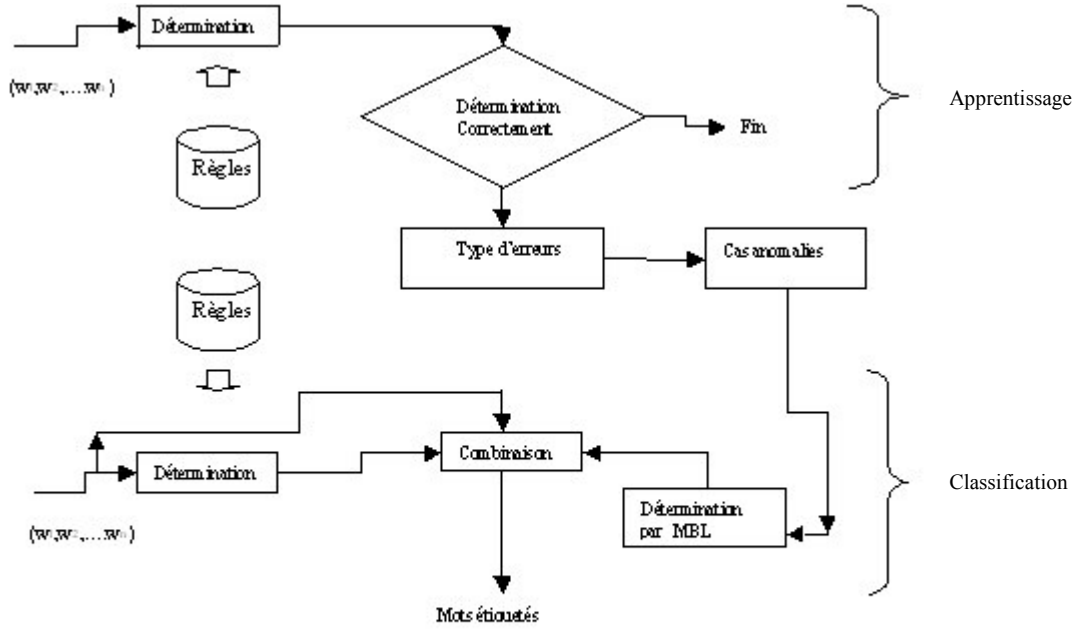

Figure 1. Architecture générale de l'étiqueteur

6 Résultats

Il suffit de comparer les différentes grammaires et dictionnaires, en particulier sur les classes de mots difficiles (adjectifs indéfinis, conjonctions, etc.), pour se convaincre de la difficulté d'établir une référence solide. La majorité des publications sur l'étiquetage automatique font l'impasse sur les cas épineux, ce qui à une conséquence directe sur l'interprétation des performances annoncées. L'application juste de l'étiquetage à base de règles donne un taux de performance de 91,87% à cause de l'ambiguïté(André, Veronis1999). Dans l'arabe les cas épineux sont nombreux, en particulier pour les noms qui ont un double rôle (nom où adjectif). Ainsi le mot de se type peut prendre une étiquette qui ne convient pas (Van Mol, 2001). Illustrons l'apport de la combinaison de la base de règles et le MBL par les exemples suivants (ces cas ambigus sont testés par le système proposé) :

Exemple 1: -جميل يشرب ici جميل nom peut prendre l'étiquette suivante: NCSgMNI.
- جوي جميل ici جميل adjective peut prendre l'étiquette suivante NACSgMNI.

Exemple 2: - دخلت بنت ici بنت un nom mais en appliquant les règles morphosyntaxiques le nom est défini comme un verbe et prend l'étiquette suivante: VPSg1 (en langue Arabe beaucoup de noms peuvent être de ce type).

Exemple 3: ما أبيض وجهه ici أبيض est un nom adjectif mais par l'application des règles morphosyntaxiques il est identifié comme un verbe et prend l'étiquette suivante : VPSg3M

Exemple 4: مدارس, أقلام, قصور. Il existe une catégorie de pluriels qui ne peut pas être identifier comme tels (A.Goweder et al.,2002). En appliquant les règles morphosyntaxiques ce type de mot peut être identifié comme singulier.

Exemple 5 : La langue arabe est très riche en particules et notre base de particules est limitée, alors une particule peut être définit comme un nom si elle ne se trouve pas dans la base de particules et ne respectant pas les règles morphosyntaxiques exemple : شتان, هيهات …etc.
Tout ces cas sont prisent par le système proposé et globalement les résultats attestent du gain apporté par le MBL, la figure 2 présente un pourcentage des cas anomalies, qui semble témoigner des limites de la base de règles(15% cas anomalies). Les résultats obtenus par la combinaison de base de règles et le MBL(figure 3) sont nettement supérieurs aux résultats obtenus par la méthode à base de règles. En particulier pour les cas de nom et de nom adjectif qui montrent que les exceptions des règles sont prises en charge par la méthode MBL. La remarque qu'on peut émettre, suite aux résultats obtenus, c'est que pour la langue arabe les informations lexicales jouent un rôle primordial dans la détermination du tag du mot en question.

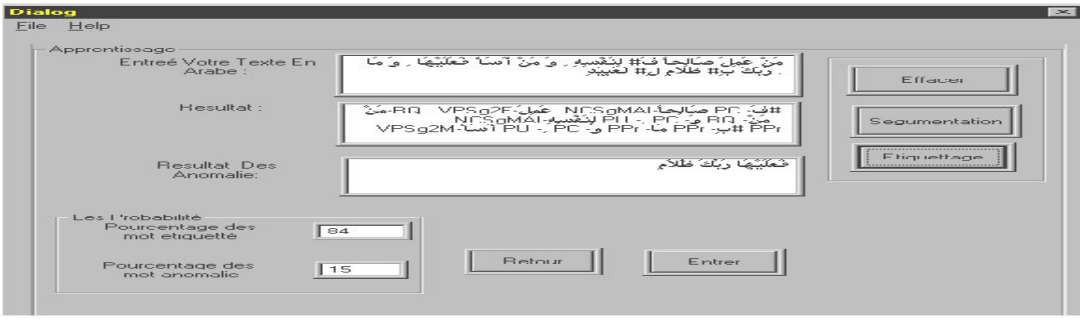

Figure 2. Résultats à base de règles

Figure 3. Résultats à base de méthode hybride

7 Conclusion

Les succès des applications de Memory-based learning dans différents domaines de recherche de bas niveau comme la reconnaissance des formes et la reconnaissance de la parole a été sans équivoque depuis l'apparition de cette approche. Néanmoins, leur utilisation pour des tâches cognitives de haut niveau telles que le traitement automatique du langage naturel a toujours suscité des remous. Dans ce contexte, le but du travail présenté dans ce papier était d'étudier la prise en charge des exceptions par les MBL plus exactement dans l'étiquetage morphosyntaxique de la langue arabe. L'étiqueteur morphosyntaxique proposé a été réalisé suite à des études approfondies sur les différentes formes d'étiquetages, qui ne sont malheureusement pas nombreux pour la langue arabe et ceux qui existe ne respectent pas les propriétés spécifique de cette langue, en appliquant à celle-ci les propriétés des langues étrangères indo-européen alors que c'est une langue sémitisée (voir MULTEXT). La comparaison nous amène à choisir l'étiquetage de sherine khoja, ce dernier est conçu spécialement pour l'arabe et l'approche hybride règle de déduction et Memory-based learning représente une performance considérable. Les perspectives envisagées pour faire évoluer le système actuel sont nombreuses. Dans un premier temps, il est possible de construire un modèle adaptatif (s'il existe des statistiques) afin de permettre le suivie de l'étiquetage et la détection avec correction automatique des erreurs. Aussi, il faut réfléchir sur l'intégration du système comme module dans un système de découpage de la phrase arabe en tronçons.

Références

Abduelbaset.,Goweder., Massimo.Poesto., Anne.De Roeck., Jeff.Reynolds.(2002); Identifying Broken Plurals in Unvowelised Arabic Text, in 2002.

Andrew.Roberts.(2003); Machine Learning in Natural language Processing, www.comp.Leeds.ac.uk.

Hacioglu K., Ward W.(2003); Target Word Detection and Semantic Role Chunking using Support Vector Machines, in *HLT-NAACL Proceedings*, pp. 25-27 , Edmonton, May 2003.

Jakub Zavrel., Walter Daelemans.(2000); Recent Advances in Memory-Based Part-of-Speech Tagging, in *Induction of Linguistic Knowledge* TSL 2000.

Mark Van Mol.(2001); The semi-automatic tagging of Arabic corpora, in *The Dutch language Union,* Amsterdam,Bulaaq, 2001.

Mona Diab., Kadri Hacioglu., Daniel Jurafsky.(2004); Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks, in *The National Science Foundation*, USA, 2004.

Saleem Abuleil., Martha Evens.(2002); Discovering Lexical Information by Tagging Arabic Newspaper Text, in *Computer and Humanities* 36(2):191-221, May 2002.

Seong-Bac Park., Byoung-Tak Zhang.(2003); Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learning, in Proceedings of the *41st Annual Meeting of the Association for Computational Linguistics,* July 2003, pp 497-504.

Shereen Khoja., Roger Garside., Gerry Knowles.(2001); A tagset for the morphosyntactic tagging of Arabic, http://www.comp.lancs.au.uk/computing/users/khoja/cl2001.pdf.

Valli André., Jean Veronis.(1999) ; Etiquetage grammatical des corpus de parole : problèmes et perspectives, http://www.up.univ-mrs.fr/~veronis/pdf/1999rfla.pdf.

Walter Daelemans., Antal van den Bosch., Jakub Zavrel., Jorn Veenstra., Sabine Buchholz., Bertjan Busser.(1998), Rapid Development of NLP Modules with Memory-based Learnig, in Proceeding of *ELSNET in Wonderland,* March 1998, pp105-113.

Walter Daelemans., Jakub.Zavrel.(1996); Part-of-Speech Tagging of Dutch with MBT, in *Informatiewetenschap* 1996, pp 33-40, The Netherlands.TU Delft.

Young-suk Lee., Kishore Papineni., Salim Roukos., Langage Model Based Arabic Word Segmentation, www.acl.ldc.upenn.edu,