Un étiqueteur sémantique des énoncés en langue arabe

Anis Zouaghi (1), Mounir Zrigui (2) et Mohamed Ben Ahmed (3)

(1) et (2) Laboratoire RIADI (unité de Monastir) - Université du centre
Faculté des Sciences de Monastir - Tunisie
(1) Anis.Zouaghi@riadi.rnu.tn
(2) Mounir.Zrigui@fsm.rnu.tn
(3) Laboratoire RIADI - Université de la Mannouba
Ecole Nationale Supérieure d'Informatique de la Mannouba - Tunisie
Mohamed.BenAhmed@riadi.rnu.tn

Mots-clefs – Keywords

Modèles statistiques de langage – Modèles n-classes – Décodage sémantique – Approche componentielle et sélective.

Statistical models of language – Models N-classes – Semantic analyze – Componential and selective approach.

Résumé – Abstract

Notre article s'intègre dans le cadre du projet intitulé Oréodule: un système de reconnaissance, de traduction et de synthèse de la parole spontanée. L'objectif de cet article est de présenter un modèle d'étiquetage probabiliste, selon une approche componentielle et sélective. Cette approche ne considère que les éléments de l'énoncé porteurs de sens. La signification de chaque mot est représentée par un ensemble de traits sémantiques Ts. Ce modèle participe au choix des Ts candidats lors du décodage sémantique d'un énoncé.

The work reported here is part of a larger research project, Oréodule, aiming at developing tools for automatic speech recognition, translation, and synthesis for the Arabic language. This article focuses on a probabilistic labelling model, according to a componential and selective approach. This approach considers only the elements of the statement carrying direction. The significance of each word is represented by a whole of semantic features Ts. This model takes part in the choice of the Ts candidates at the time of the semantic decoding of a statement.

1 Introduction

Depuis quelques années, La tendance est vers l'utilisation des modèles de langages statistiques dans le domaine de la compréhension automatique de la parole spontanée (Bousquet, 2002), (Lefèvre, 2002), etc. Pour la langue arabe, l'utilisation de tels modèles à notre connaissance constitue une nouveauté. L'avantage principal de ces modèles statistiques par rapport aux modèles à syntaxe fixe (Bennacef et al., 1994) est qu'ils sont plus portables vers d'autres domaines (Minker, 1999), et nécessite moins de recours à un expert humain. Dans cet article, nous proposons un étiqueteur sémantique basé sur un modèle de langage probabiliste hybride [Zouaghi et al., 2005] pour l'interprétation d'une séquence de mots reconnue par le module de reconnaissance de la parole. Ce modèle participe au choix des ensembles de traits sémantiques Ts candidats, en tenant compte des données suivantes: le type d'acte illocutoire accompli par l'énoncé (demande, refus, excuse, etc.), le type de l'énoncé (demande de réservation, de tarifs, etc.), des mots déjà interprétés (les traits sémantiques utilisés), et de la probabilité d'interprétation d'un mot par un Ts candidat.

2 Modèle probabiliste

2.1 Corpus d'apprentissage

Le corpus d'apprentissage considéré décrit des demandes de renseignements ferroviaires, en langue arabe classique. Chaque mot significatif dans ce corpus se voit attribuer un ensemble de traits (Ts), tel que défini dans (Zouaghi et al., 2004). Le mot الذاهب (qui va) par exemple se voit attribuer Ts = (Transport_Ferroviaire, Mouvement, Destination). Les mots synonymiques ou possédant un même rôle sémantique sont interprétés via un même Ts. Pareil, pour les mots dérivés à partir d'une même racine morphologique et possédant un même sens (tels que الذاهب (qui va) et يذهب (va) qui sont dérivés à partir de la racine ذهب (dhahaba)). Nous avons utilisé une quarantaine de Ts différents pour l'étiquetage du corpus. En plus chaque énoncé de ce corpus se voit attribuer une étiquette permettant de préciser le type de l'énoncé. En tout, nous avons utilisé sept étiquettes.

Domaine	Taille (Mo)	Nombre d'énoncés	Nombre de mots	Nombre de locuteurs
Renseignements ferroviaires	*3,4*	*10000*	*85900*	*1000*

Figure 1 : Caractéristiques du corpus de point de vue de son volume.

Nature de la tâche	Renseignements sur les:				Réservations	autres
Taux de sa représentation	horaires	trajets	tarifs	durées	*10,41 %*	*40,64%*
	28,7 %	*9,37 %*	*16,66 %*	*3,12 %*		

Figure 2 : Caractéristiques du corpus de point de vue de son contenu.

Ce corpus a été collecté en demandant à cent personnes de formuler des énoncés relatifs aux renseignements ferroviaires. Donc c'est un corpus simulé et non pas réel. L'inconvénient de ce type de corpus est qu'il ne permet pas de décrire parfaitement l'application.

2.2 Principe du décodage sémantique

Nous entendons par décodage sémantique d'un énoncé, l'étiquetage de chacun de ses mots significatifs via un Ts (Zouaghi et al., 2004). Seulement les mots porteurs de sens parmi ceux qui sont reconnus sont interprétés.

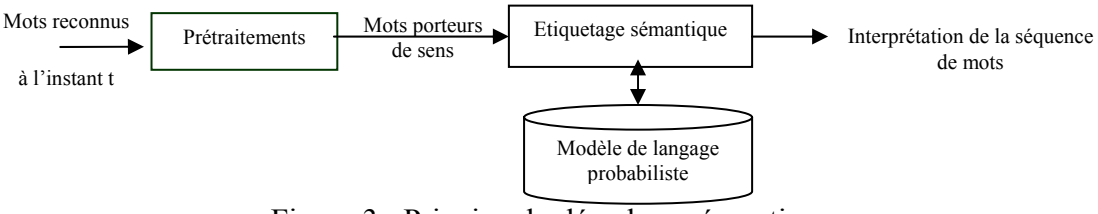

Figure 3 : Principe du décodage sémantique.

Soit la séquence de mots significatifs S = M1 M2 M3 M4 obtenus après la phase de prétraitement (figure 3). Soit Ts1, Ts2 et Ts3 les traits affectés respectivement aux mots M1, M2 et M3. A partir de ces données, nous voulons déterminer le Ts correspondant à M4. Pour atteindre cet objectif, nous utilisons un modèle de langage probabiliste hybride, permettant de tenir compte du type et de la nature de l'énoncé, ainsi que des mots déjà interprétés (figure 4).

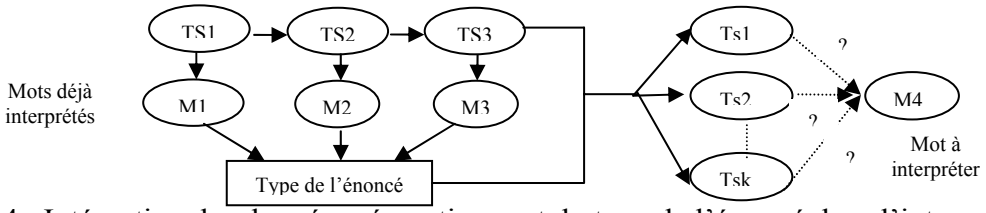

Figure 4 : Intégration des données sémantiques et du type de l'énoncé dans l'interprétation.

2.3 Description du modèle

Les systèmes à base de modèles de langage probabilistes tentent de déterminer le score d'une séquence de mots $S = m_1, m_2, ..., m_i$, dont la formule générale est la suivante:

$$P(S) = P(m_1).P(m_2/m_1) \dots P(m_i/m_1, m_2, \dots, m_{i-1}) \quad (1)$$

Dans le cas de l'étiquetage I d'une séquence de mots significatifs $M_1 \dots M_n$, par $Ts_1 \dots Ts_n$, le modèle tente de déterminer le score d'interprétation de chacun de ces mots, par chacun de ces traits. Soit $I=Ts_1 \rightarrow M_1 \dots Ts_n \rightarrow M_n$, la vraisemblance de I est alors définie comme suit:

$$P(I) = P(Ts_1 \dots Ts_n | M_1 \dots M_n) = P(Ts_1/M_1 \dots M_k).\ P(Ts_2/Ts_1, M_1 \dots M_k) \dots P(Ts_n/Ts_1 \dots Ts_{n-1}, M_1 \dots M_n) = P(Ts_1/M_1)\ .\ P(Ts_2/Ts_1, M_2) \dots .P(Ts_n / Ts_1 \dots Ts_{n-1}, M_n) \quad (2)$$

Nous signalons que le passage de la deuxième à la troisième ligne correspond à une approximation du modèle, qui considère que la probabilité d'un Tsi ne dépend, conditionnellement à la séquence complète des traits, qu'au mot courant Mi. En fixant à l'avance le domaine de l'application, chaque mot significatif M_i peut être interprété via $Ts_i = (C_i, TM_i)$, où Ci indique la classe à laquelle appartient le mot M_i, et TM_i le trait micro sémantique qui lui correspond. L'équation (2) devient:

$$P(I)=P((C_1,TM_1)/M_1).P((C_2,TM_2)/(C_1,TM_1),M_2)..P((C_n,TM_n)/(C_1,TM_1)\dots(C_{n-1},TM_{n-1}),M_n) \quad (3)$$

Nous avons intégré dans l'équation (4) d'autres sources d'informations afin d'améliorer la qualité du décodeur sémantique. Ceci, en tenant compte du type de l'énoncé noté par NT_j (avec P ($NT_j/M_1 \ldots M_n$) est la probabilité conditionnelle d'avoir un énoncé de type NT_j).
$P(I) = P(Ts_1 \ldots Ts_n | NT_j, M_1 \ldots M_n) = P(NT_j/M_1 \ldots M_n).\ P((C_1, TM_1)/NT_j, M_1) \ldots P((C_n, TM_n)/NT_j, (C_1, TM_1) \ldots (C_{n-1}, TM_{n-1}), M_n)$ *(4)*

2.4 Lissage du modèle

La première approximation appliquée à ce modèle consiste à ne considérer pour la détermination du type de l'énoncé, que certains mots appelés *mots de référence* notés Mr_k. Les mots de référence sont des mots dont leurs occurrences dans un énoncé permettent de déterminer son type. Ces mots sont en fait des uni-grammes, ou des bi-grammes, et dans certains cas des tri-grammes, dont la probabilité est égale à un. Par exemple le bi-gramme القطار (le train) أريد (je veux) constitue un mot de référence, permettant d'identifier les énoncés de type réservation. Ce bi-gramme ne peut être rencontré que dans les énoncés du corpus d'apprentissage qui sont étiquetés par l'étiquette <Réservation> (on a: P (القطار (le train)/أريد (je veux))=1). On obtient ainsi la substitution suivante:
$P(NT_j / M_1 \ldots M_n) = P(NT_j / Mr_k)$ *(5)*
La deuxième hypothèse de modélisation porte sur les relations d'indépendance conditionnelle dans le modèle et concerne la probabilité jointe $P((C_i,TM_i)/NT_j,(C_1,TM_1)\ldots(C_{i-1},TM_{i-1}), M_i)$.
$P((C_i,TM_i)/NT_j,(C_1,TM_1)\ldots(C_{i-1},TM_{i-1}),M_i) =$
$P(C_i / NT_j, (C_1, TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i).\ P(TM_i / Ci, NT_j, (C_1,TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i)$ *(6)*
Afin de simplifier ce modèle, nous avons considéré seulement les Ts jugés pertinents TsP (CP, TMP) à la prédiction du Ts correspondant au mot M_i noté par $Ts(M_i)$. Un Ts n'est considéré pertinent, que lorsqu'il est suivi par un nombre k minime de Ts (k tend vers 1). Nous avons fixé k = 3, car nous pensons que pour k= 1, la grammaire devient très rigide et ça revient à considérer dans l'historique du mot M_i que les mots jouant le rôle de marqueurs (Fillmore, 1968). Par exemple, l'ensemble de traits Ts = (مؤشر_حركة (Indice_mouvement), مؤشر_وجهة (Indice_destination)) est un Ts pertinent car cet ensemble est toujours succédé dans le corpus d'apprentissage par Ts = (مدينة (ville), وجهة (destination)). k correspondant à Ts=(Indice_mouvement, Indice_destination) est ainsi égal à 1. La deuxième approximation considérée est que C_i à un instant t, ne dépend que des classes pertinentes précédentes CP et du type de l'énoncé, on a ainsi:
$P(C_i / NT_j, (C_1, TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i) = P(C_i/NT_j, CP_{i-1}, \ldots, CP_{i-l})$ *(7)*
Une autre approximation considérée, est que TM_i du $Ts(M_i)$, à un instant t, ne dépend que de la classe C_i affectée à M_i et du trait pertinent précédent $TsP_{i-1}(CP_{i-1}, TMP_{i-1})$. Ainsi on a:
$P(TM_i / Ci, NT_j, (C_1, TM_1) \ldots (C_{i-1}, TM_{i-1}), M_i) = P(TM_i / Ci, CP_{i-1}, TMP_{i-1})$ *(8)*
A partir de ces deux approximations (7) et (8), on déduit de l'équation (6) que:
$P((C_i,TM_i)/NT_j,(C_1,TM1)..\ (C_{i-1},TM_{i-1}),M_i) = P(C_i/NT_j,CP_{i-1}..\ CP_{i-l}).P(TM_i/Ci,CP_{i-1},TMP_{i-1})$ *(9)*
Et enfin à partir des équations (5) et (9), on déduit à partir de l'équation (4) que:
$P((C_i,TM_i) \rightarrow M_i/NT_j) = P((C_i,TM_i)/M_i,NT_j)$
$= P(N_j / Mr_k).\ P(C_i/NT_j, CP_{i-1}, \ldots, CP_{i-l}) \ldots P(TM_i / Ci, CP_{i-1}, TMP_{i-1})$ *(10)*

2.5 Applicabilité du modèle à la langue arabe

Comme signalé ci-dessus, nous avons considéré une approche sélective (voir paragraphe 2.2).

Dû aux spécificités de la langue arabe, on peut s'interroger sur l'adéquation de cette approche au traitement de cette langue. L'absence de la voyellation est l'une des sources d'ambiguïtés majeure de la compréhension de cette langue. Pour mieux comprendre, le mot non voyellé ذهب (thalhab) par exemple peut avoir deux significations différentes selon la manière de sa voyellation. Il a le sens du verbe partir en le prononçant (thahaba), et de l'or lorsqu'il est prononcé (thahabon). Ce mot peut être ainsi interprété par Ts= (حركة (Mouvement), وجهة (destination)), ou par Ts=((métal) معدن , ثمين (cher)). Or la détermination de la voyellation correspondante à un mot (et par conséquent son sens), nécessite plusieurs niveaux de connaissances: morphologiques, syntaxiques, ... (Debili et al., 2002). Cette nécessité est surmontée dans notre cas par la nature du domaine restreint de l'application. Nous prospectons d'améliorer la performance de l'étiqueteur, en lui intégrant des données syntaxiques.

3 Application du modèle

Nous avons utilisé une centaine d'énoncés (différents du ceux du corpus d'apprentissage), portant tous sur des demandes d'horaires pour le test. Le corpus d'apprentissage a été étiqueté avec 37 Ts. Pour juger de la qualité de notre étiqueteur, nous avons calculé le pourcentage d'étiquettes sémantiques qui sont incorrectement attribuées, à partir de la formule suivante: $Taux\_erreur=N_{inc}/N\times100$. Où, N_{inc} est le nombre de Ts incorrectement attribués, et N est le nombre total des Ts attribués par un expert au corpus de test. N est égal à 500 dans ce test. La table ci-dessous montre les Taux_erreur des étiqueteurs sémantiques obtenus en considérant des modèles bi-classes et tri-classes ainsi que le modèle hybride défini. La longueur de l'historique est fixée à 3 pour la détermination des C_i et à 2 pour TM_i.

Etiqueteurs sémantiques considérés	Taux_erreur
bi-classes: (1) $P((C_i, TM_i) / M_i) = P(C_i / C_{i-1})) \times P(TM_i / Ci, Ts_{i-1})$	57%
avec considération lexicale: (2) $P((C_i, TM_i) / M_i) = P(C_i / M_{i-1}, C_{i-1})) \times P(TM_i / M_i, Ci, Ts_{i-1})$	45%
tri-classes: (1) $P((C_i, TM_i) / M_i) = P(C_i / C_{i-1}, C_{i-2})) \times P(TM_i / Ci, Ts_{i-1})$	48,6%
(2) $P((C_i, TM_i) / M_i) = P(C_i / M_{i-1}, C_{i-1}, C_{i-2})) \times P(TM_i / M_i, Ci, Ts_{i-1})$	41,2%
hybride k=2: (1) $P((C_i, TM_i)/M_i, NT_j) = P(NT_j) \times P(C_i/NT_j, CP_{i-1}, CP_{i-2})) \times P(TM_i / Ci, TsP_{i-1},)$	50%
(2) $P((C_i, TM_i)/M_i, NT_j) = P(NT_j) \times P(C_i /NT_j, M_{i-1},CP_{i-1},CP_{i-2})) \times P(TM_i / M_i, Ci, TsP_{i-1})$	39,4
hybride k=3: (1) $P((C_i, TM_i) / M_i, NT_j) = P(NT_j) \times P(C_i / NT_j, CP_{i-1}, CP_{i-2})) \times P(TM_i / Ci,TsP_{i-1})$	46,8
(2) $P((C_i, TM_i)/ M_i, NT_j) =P(NT_j) \times P(C_i / NT_j, M_{i-1}, CP_{i-1}, CP_{i-2}))\ \times P(TM_i /M_i,Ci, TsP_{i-1})$	37%

Figure 5 : Taux d'erreur des étiqueteurs sémantiques considérés.

4 Interprétation des résultats

D'après la table ci-dessus, chaque fois que l'on intègre des données lexicales dans un modèle, le résultat s'améliore. Nous avons utilisé l'approche de (Katz, 1987) pour l'estimation des données manquantes. L'amélioration est encore meilleure, en considérant en même temps le type de l'énoncé et les Ts pertinents, pour la prédiction du Ts suivant. Nous remarquons que malgré l'amélioration de la qualité de l'étiqueteur sémantique, le taux d'erreur (qui atteint 37%) reste comme même un peu élevé. Ceci est dû au fait, que certains énoncés du corpus de test ont une structure syntaxique très complexe. Afin de remédier ce problème, certains

systèmes combinent une analyse syntaxique profonde avec une analyse sélective tel que le système TINA de (Seneff, 1992). D'autres systèmes utilisent les stratégies d'analyses du TAL robuste (Antoine et al., 2003). Ces systèmes sont performants dans des applications ouvertes.

5 Conclusion

Nous avons présenté dans cet article un étiqueteur sémantique basé sur un modèle de langage hybride. Ce modèle permet d'intégrer des données contextuelles lexicales, sémantiques ainsi qu'illocutoire en même temps. Il permet en plus de ne tenir compte que des traits sémantiques pertinents dans l'historique du mot à interpréter. Afin de montrer l'avantage de ce modèle, nous l'avons évalué et comparé par rapport aux modèles n-classes classiques, qui ne tiennent pas compte de la nature et du type de l'énoncé dans le calcul de la probabilité d'interprétation d'un mot par un Ts donné.

Références

Antoine J-Y., Goulian J., Villaneau J. (2003), Quand le TAL robuste s'attaque au langage parlé: analyse incrémentale pour la compréhension de la parole spontanée, Actes de *TALN*.

Bennacef S., Bonneau-Maynard H., Gauvain J-L., Lamel L., Minker W. (1994), A spoken language for information retrieval, Actes de *ICSLP*, 1271-1274.

Bousquet-Vernhettes C. (2002), C*ompréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique*, Thèse de l'université de Toulouse III, 84-85.

Débili F., Achour H., Souici E. (2002), La langue arabe et l'ordinateur: de l'étiquetage grammatical à la voyellation automatique, *Correspondances de l'IRMC,* N° 71, 10-28.

Fillmore C. J. (1968), *The case for case,* Hollt and Rinehart and Winston Inc.

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

Lefèvre F. (2000), *Estimation de probabilité non paramétrique pour la reconnaissance markovienne de la parole,* Thèse de l'Université Pierre et Marie Curie.

Minker W. (1999), *Compréhension automatique de la parole spontanée*, Paris, L'Harmattan.

Seneff S. (1992), Robust parsing for spoken language systems, Actes de *ICASSP*, 189-192.

Zouaghi A., Zrigui M., Ben Ahmed M. (2004), Une structure sémantique pour l'interprétation des énoncés en langue arabe, Actes de *JEP-TALN-ARABIC*.

Zouaghi A., Zrigui M., Ben Ahmed M. (2005), A statistical model for semantic decoding of Arabic language statements, Actes de *NODALIDA*.