Calcul du sens des mots arabes ambigus

Anis Zouaghi[1], Mounir Zrigui[1], Mohamed Ben Ahmed[2]

[1] Unité de Monastir – Labo RIADI
anis.zouaghi@riadi.rnu.tn ; mounir.zrigui@fsm.rnu.tn

[2] Université de la Mannouba – Labo RIADI
mohamed.benahmed@riadi.rnu.tn

Résumé

Nous présentons dans cet article un analyseur sémantique pour la langue arabe. Cet analyseur contribue à la sélection du sens adéquat parmi l'ensemble des sens possibles que peut recevoir un mot hors contexte. Pour atteindre cet objectif, nous proposons un modèle vectoriel qui permet de lever les ambiguïtés locales au niveau de la phrase et celles relevant du domaine. Ce modèle est inspiré des modèles vectoriels très utilisés dans le domaine de la recherche documentaire.

Mots-clef : désambiguïsation sémantique, modèle vectoriel, traitement de la parole arabe, influence sémantique.

Abstract

This article describes a semantic analyzer for the Arabic language. This analyzer contributes to the selection of the adequate meaning among the set of possible meanings for a given word. To achieve this goal, we propose a vectorial model that allows lifting local ambiguities on the level of the sentence and those concerning semantic domains. This model is inspired from vector models commonly used in information retrieval.

Keywords: semantic disambiguation, vector models, processing of Arabic speech, pertinent context, semantic influence.

1. Introduction

Notre travail s'intègre dans le cadre du projet Oréodule : un système de reconnaissance, de traduction et de synthèse de la parole spontanée. L'objectif de cet article est de présenter un analyseur sémantique des mots arabes ambigus. Contrairement à la plupart des analyseurs utilisés dans les systèmes de compréhension de la parole, basés sur les modèles HMM (Minker, 1999 ; Bousquet, 2002), notre analyseur est basé sur un modèle vectoriel. Ce modèle permet de représenter chaque sens possible par un vecteur sémantique, composé des mots ayant une affinité sémantique avec le mot ambigu. Ce modèle est inspiré des modèles vectoriels utilisés dans le domaine de la recherche documentaire. Bien que la désambiguïsation sémantique possède un enjeu important dans les applications TALN (Ide *et al.*, 1998), les ressources nécessaires pour résoudre ce problème restent presque indisponibles pour la langue arabe. Ceci nous a amené à créer notre propre corpus d'entraînement, et de l'étiqueter sémantiquement en attribuant à chaque mot ambigu l'ensemble de traits sémantiques Tse approprié, où chaque Tse est constitué de 3 traits et représenté comme suit Tse = (domaine, classe sémantique, trait micro sémantique).

2. Méthode de calcul du sens

L'interprétation d'un mot ambigu est obtenue suite à la coopération de 2 étapes d'analyse. La 1re étape correspond à la levée des ambiguïtés relevant du domaine. Elle permet de déterminer les ensembles Tse représentant une probabilité faible pour décrire le sens du mot ambigu dans le texte où il est apparu, alors que la 2e étape correspond à une analyse plus fine. Cette analyse est basée sur l'étude des affinités sémantiques entre le mot ambigu et les mots qui l'entourent dans l'énoncé.

3. Étape de levée des ambiguïtés relevant du domaine

Au cours de cette étape d'analyse, nous considérons l'influence du domaine sur la caractérisation du sens d'un mot. Ainsi, notre modèle considère une fenêtre d'analyse de taille assez grande (égale à la longueur du texte). À partir de l'inventaire des sens possibles du mot ambigu, est calculée la probabilité d'interprétation du mot MA avec chaque Tse_i possible. Cette probabilité est calculée comme suit :

$$P_{ARD}(Tse_i / MA) = P(D_k) \times P(Tse_i / MA, D_k) \quad (1)$$

L'équation (1) tient compte de l'influence du domaine dans le calcul du sens de MA. Ceci par la considération des 2 probabilités $P(D_k)$ et $P(Tse_i / MA, D_k)$. $P(D_k)$ calcule la probabilité que le texte auquel appartient MA appartient au domaine D_k ; $P(Tse_i / MA, D_k)$ est la probabilité que le sens Tse_i soit affecté au mot MA sachant que le texte appartient au domaine D_k.

3.1. Calcul du domaine décrit par un texte

L'identification du domaine auquel appartient le mot ambigu est obtenue à partir de la probabilité $P(D_k)$, en se basant sur un modèle vectoriel. Ce modèle permet de caractériser chaque domaine par un ensemble de mots-clés. À chaque mot-clé est attribué un poids p_{ij} en utilisant la méthode Tf-Idf (Term frequency – Inverse document frequency). Ainsi le poids p_{ij} d'un mot m_i dans un texte décrivant un domaine D_j est obtenu à partir de l'équation suivante :

$$p_{ij} = [tf(m_i, D_j) \times \log(n / df(m_i))] / [tf(m_i, D_j) + 0.5 + (1.5 \times n \times l(D_j) / \sum_{Dk} l(Dk)) \times \log(n+1)] \quad (2)$$

où n et $l(D_k)$ désignent respectivement le nombre des domaines considérés et la longueur de l'ensemble des textes représentant le domaine D_k ; le terme $tf(m_i, D_j)$ désigne le nombre d'occurrences de m_i dans D_j ; $df(m_i)$ correspond au nombre de domaines où apparaît m_i.

À partir de ces poids, est associé à chaque domaine D_j considéré un vecteur caractéristique présenté comme suit : $\vec{D_j}=(p_{1j}, p_{2j}, p_{3j},.., p_{ij},\ldots, p_{nj})$; avec $1 \leq n \leq l$, où l est le nombre total de tous les mots considérés pertinents pour l'identification de chacun des domaines considérés.

Nous définissons la probabilité notée $P(D_j^T)$ qu'un texte T décrit un domaine D_j dont le vecteur caractéristique est $\vec{D_j} = (p_{1j}, p_{2j}, p_{3j},.., p_{ij},\ldots, p_{nj})$, comme la somme des probabilités p_{ij} des mots pertinents rencontrés dans le texte T :

$$P(D_j) = P(D_j^T) = \sum_{mi \in T} pij \quad (3)$$

3.2. Calcul du sens en considérant l'influence du domaine

Comme signalé ci-dessus, notre modèle calcule le sens d'un mot ambigu MA en tenant compte de l'influence sémantique du domaine sur celui-ci. Ceci est réalisé en utilisant la probabilité conjointe $P(Tse_i / MA, D_k)$ dont la formule est donnée par l'équation (4) suivante :

$$P(Tse_i / MA, D_k) = N(Tsei(MA), D_k) / N(MA, D_k) \quad (4)$$

où $Tse_i(MA)$ est l'instanciation du sens de MA par l'ensemble Tse_i, et le terme N ($Tse_i(MA)$, D_k) désigne le nombre d'interprétation de MA par Tse_i dans le domaine D_k. N(MA, D_k) est en faite tf (MA, D_k) qui désigne le nombre d'occurrences de MA dans D_k.

4. Étape de levée des ambiguïtés locales

Cette étape permet de calculer le sens adéquat du mot ambigu en se basant sur une analyse sémantique locale. Pour cela, nous considérons une fenêtre d'analyse plus réduite que celle considérée dans la première étape lors de la levée des ambiguïtés relevant du domaine. La taille de cette fenêtre est égale à la longueur du contexte droit du mot cible dans la phrase où il est apparu. Pour lever les ambiguïtés locales, nous représentons chaque sens possible d'un mot ambigu MA par un vecteur sémantique noté <MA, Tse_j>. Ce vecteur permet de caractériser chaque sens possible Tse_j de MA par les mots qui ont une influence sémantique sur MA. À chacun de ces mots est attribué un poids q_{ij} déterminé à partir de la formule (5) suivante :

$$q_{ij} = N(m_i, Tse_j(MA) / N(m_i) \quad (5)$$

où le terme N (m_i, Tse_j(MA)) désigne le nombre de cooccurrence du mot m_i avec MA dans une même phrase, sachant que ce dernier est instancié par le sens Tse_j, et N(m_i) désigne le nombre total d'occurrence de m_i avec MA dans un même contexte (phrase). Ce poids q_{ij} peut être aussi obtenu en utilisant une des méthodes de mesure de la similarité sémantique entre deux mots, telle que l'information mutuelle moyenne (Rosenfeld, 1994). À partir de ces poids q_{ij} est associé à chaque sens possible de MA un vecteur sémantique caractéristique présenté comme suit : <MA, Tse_j> = (q_{1j}, $q_{2j, ..,}$ $q_{kj, ...,}$ q_{nj}). Les mots considérés comme ayant une influence sémantique sur MA sont les mots dont le poids q_{ij} est supérieur à un seuil donné. Pour identifier le sens Tse_j adéquat à attribuer au mot MA dans une phrase P, nous utilisons le modèle P_{AL} décrit comme suit :

$$P_{AL}(Tse_j/MA) = \sum_{mi \in Cd} qij \quad (6)$$

Ce modèle définit la probabilité d'affecter le sens Tse_j au mot MA, et dont le vecteur sémantique est <MA, Tse_j> = (q_{1j}, $q_{2j, ..,}$ $q_{kj, ...,}$ q_{nj}), comme la somme des probabilités q_{ij} des mots m_i influant sur le sens de MA, rencontrés dans le contexte droit Cd du mot MA.

5. Analyseur sémantique des mots ambigus

Notre modèle calcule le sens d'un mot en tenant compte des ambiguïtés locales et des ambiguïtés relevant du domaine. Nous avons ainsi combiné les 2 modèles P_{ARD} et P_{AL}, à partir de l'équation suivante :

$$P(Tse_i/MA) = \lambda \times P_{ARD}(Tse_i/MA) + \rho \times P_{AL}(Tse_i/MA) \quad (7)$$

avec λ et ρ deux coefficients à déterminer empiriquement à travers des tests et des comparaisons de pertinence.

6. Application du modèle

Pour l'évaluation de notre modèle, nous avons utilisé 100 énoncés (859 mots) décrivant des demandes de renseignements ferroviaires en langue arabe. Nous avons testé chacun des modèles définis séparément, afin de pouvoir juger de leur efficacité et étudier leurs limites. En ce qui concerne le modèle P_{AL}, nous l'avons appliqué pour déterminer le rôle sémantique accompli par une ville (ville de départ, ou d'arrivée, ou de stop ou de correspondance). Pour

évaluer le modèle P_{ARD}, nous l'avons appliqué pour la détermination du type et de l'acte illocutoire accompli par un énoncé.

6.1. Évaluation du modèle de calcul du domaine

Comme mentionné dans le paragraphe 3.1, chaque domaine est caractérisé par un ensemble de mots-clés. Dans notre cas il s'agit d'attribuer à chaque type d'énoncé, un vecteur caractéristique $\overrightarrow{D_j}$. Pour cela, nous avons utilisé 5 vecteurs pour la représentation des 5 types d'énoncés considérés (voir tableau 1). Pour réduire la dimension de ces vecteurs, les mots-clés dérivés d'une même racine sont regroupés en une seule entrée. Par exemple, les mots-clés احجز, الحجز, حجز identifiant les énoncés de type réservation sont transformés en حجز (réserver). Les ensembles de termes {توقيت, أوقات الأوقات, بأوقات}, {أسعار, سعر}, {العبور, يعبر}, { مدة, } et {ثمن, الثمن}, {المدة} sont regroupés en وقت, عبر, مد et ثمن.

Type de l'énoncé	Mots-clés	Vecteurs caractéristiques (les valeurs sont obtenues en utilisant l'équat. 2)
Demande d'horaire	بأوقات, الأوقات, أوقات, توقيت, متى, ساعة, أي, الأيام	(0.274; 0; 0; 0; 0; 0; 0; 0; 0; 0.351; 0.142; 0.351; 1.111; 0)
Demande Tarif	بكم, سعر, أسعار, الثمن, ثمن, تعريفة	(0; 0; 0; 0,669; 0; 0; 0; 1.226; 0; 0; 0; 0; 0.577)
Demande de durée	مدة, المدة, يمضي, يمكث	(0; 0; 0,27; 0; 0; 0,832; 0,832; 0; 0; 0; 0; 0; 0)
Demande de trajet	يعبر, العبور	(0; 0,661; 0,075; 0; 0; 0; 0; 0; 0; 0.075; 0; 0; 0)
Demande de réservation	حجز, الحجز, احجز	(0.1; 0; 0; 0; 1.243; 0; 0; 0; 0; 0; 0; 0; 0)

Tableau 1. Vecteurs caractéristiques $\overrightarrow{D_j}$ associés à chaque type d'énoncé

Ainsi chaque vecteur caractéristique est de la forme $\overrightarrow{D_j}$ = (بكم, يمضي, يمكث, حجز, ثمن, مد, عبر, وقت, تعريفة, متى, ساعة, أي, الأيام). En appliquant par exemple ce modèle au calcul du type de l'énoncé (E) suivant : أريد حجز 4 تذاكر من الصنف الأول في القطار لسفرة بين سوسة و بنزرت Je veux réserver 4 billets de classe première dans le train pour un voyage entre Sousse et Bizerte). On trouve que cet énoncé est de type demande de réservation ($P(\overrightarrow{DRéservation}^E) = 1,243$). Ce modèle est efficace face aux énoncés ne contenant pas des autocorrections. Pour améliorer les performances du modèle face aux autocorrections, nous avons introduit dans les vecteurs $\overrightarrow{D_j}$ des classes de mots C^t_{ij}. Une classe C^t_{ij} est appelée modificateur du poids du mot i dans les énoncés de type $\overrightarrow{D_j}$, en utilisant un contexte de taille t. On a remplacé par exemple dans $\overrightarrow{D}_{réservation}$ حجز « réserver » par ($C^3_{4,réservation}$ + حجز), où $C^3_{4,réservation}$={بل « plutôt »}. Ainsi, si on rencontre dans le contexte gauche (de taille 3) du terme حجز la classe $C^3_{4,\ réservation}$, le poids $p_{4,réservation}$ associé à حجز dans le vecteur $\overrightarrow{D}_{réservation}$ sera modifié en utilisant la formule suivante : $p_{ij} = p_{ij} - m\ (C^t_{ij})$. Où $m\ (C^t_{ij})$ est la masse associée à la classe C^t_{ij}. Pour la classe $C^3_{4,réservation}$, la masse $m(C^3_{4,\ réservation}) = - p_{4,réservation}$. En utilisant ce modèle vectoriel modifié le taux d'erreur est passé de 90 % à 5 %. Soit par exemple l'énoncé (E1) suivant :

Je veux réserver non plutôt connaître l'horaire d'allée ← أريد حجز لا بل معرفة وقت الذهاب
Comme le montre la figure 1, le modèle vectoriel modifié est plus efficace. L'énoncé (E1) est identifié de type demande d'horaire grâce à la présence du terme بل « plutôt », qui active la classe $C^3_{4,\ réservation}$ qui modifie le poids $p_{4,\ réservation}$ correspondant à حجز de $\overrightarrow{D}_{réservation}$. La taille « 3 » du contexte à considérer pour l'activation de $C^3_{4,\ réservation}$ est déterminée empiriquement.

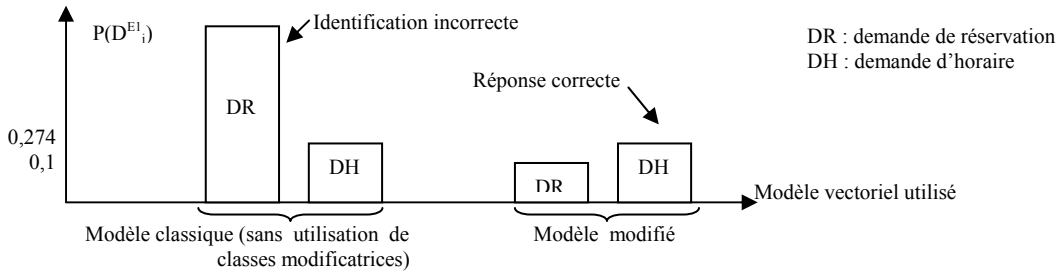

Figure 1. Résultats obtenus via des modèles vectoriels classique et modifié

6.2. Évaluation du modèle de levée des ambiguïtés locales

Nous avons évalué le modèle P_{AL} en l'appliquant pour l'interprétation d'une ville. Afin de simplifier le modèle, nous avons regroupé toutes les villes dans une même classe VILLE. Nous avons utilisé 4 vecteurs <VILLE,départ>, <VILLE,arrivée>, <VILLE,stop> et <VILLE,correspondance>, pour représenter les sens possibles de chaque ville. Le tableau 2 représente le vecteur caractéristique correspondant à chaque sens. Chaque vecteur est représenté comme suit : (إلى « à », نحو « vers », اتجاه « direction », نزل « descendre », وقف « arrêter », عبر « passer », يمر « traverse », من « de », ينطلق « part », ذهب « aller », بين « entre »). Les mots dérivés d'une même racine sont regroupés, tels que عبر = {عبر « à travers », عبر « passe »}, وقف = {يتوقف , يقف}, نزل = {انزل, النزول} et ذهب = {الذاهب , يذهب}.

Rôles sémantiques	Vecteurs sémantiques caractéristiques
Ville de départ	(0; 0; 0; 0; 0; 0; 0; 0.8; 0.33; 0; 1)
Ville d'arrivée	(1; 1; 1; 1; 0; 0; 0; 0; 0.66; 1; 0)
Ville de stop	(0; 0; 0; 0; 1; 0; 0; 0; 0; 0; 0)
Ville de correspondance	(0; 0; 0; 0; 0; 1; 1; 0.2; 0; 0; 0)

Tableau 2. Vecteurs caractéristiques correspondant aux interprétations d'une ville.

Ci-dessous un exemple d'interprétation de deux villes occurrent dans un même énoncé en utilisant le modèle P_{AL} : هل يمكنك إعلامي بكم سعر التذكرة للذهاب من القلعة إلى تونس → Est-ce que tu peux m'informer, combien le prix du billet pour aller de Kalaâ à Tunis.

L'interprétation de القلعة « Kalaâ » comme ville de départ est réalisée grâce à la présence du terme من « de » (dont le poids est égal à 0.8) dans le contexte droit de celui-ci. Alors que la ville تونس « Tunis » est interprétée comme étant une ville d'arrivée, grâce à la présence de إلى « à » dans le contexte se trouvant entre les deux villes (poids correspondant est égal à 1). L'avantage de notre modèle est que les contraintes influant sur le sens d'un mot sont déterminées automatiquement, sans avoir besoin d'une grammaire à base de règles comme dans (Bennacef, 1994). Les mauvaises interprétations sont dues à la non-présence dans l'énoncé des termes appartenant aux vecteurs caractéristiques des sens possibles d'une ville.

6.3. Évaluation du modèle de calcul du domaine

La figure 2 présente les résultats d'interprétation (Zouaghi *et al.*, 2005), obtenus par l'application de notre modèle et des modèles n-classes classiques (pour n=2 et n=3). Les modèles n-classes permettent d'attribuer un sens Tse_i à un mot en fonction d'un historique h, en utilisant la formule suivante : $P(Tse_i / h) = P(MA / Tse_i) \times P(Tse_i / h)$.

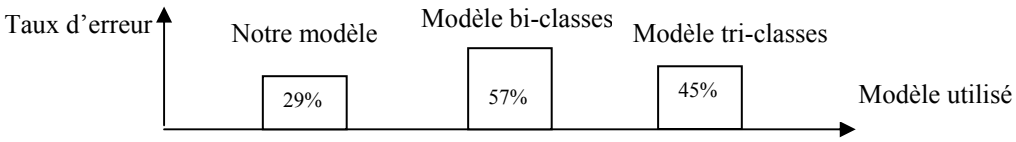

Figure 2. Résultats obtenus

On remarque que le meilleur score d'interprétation (71 %) est obtenu en utilisant notre modèle vectoriel.

7. Conclusion

Nous avons présenté un modèle pour le calcul du sens des mots arabe ambigus dans le domaine de la compréhension de la parole. Ce modèle est inspiré des modèles vectoriels utilisés dans le domaine de la recherche documentaire. Il contribue à la sélection du sens adéquat, à la suite de la coopération de deux étapes d'analyse. Une analyse permettant de lever les ambiguïtés relevant du domaine, et une autre pour lever les ambiguïtés locales au niveau de la phrase. Afin d'être robuste face aux autocorrections, nous avons introduit des classes à l'intérieur des vecteurs, afin de corriger les poids du vecteur en cas d'autocorrection. Les tests effectués prouvent la bonne performance de notre modèle.

Références

BENNACEF S., BONNEAU-MAYNARD H., GAUVAIN J-L., LAMEL L., MINKER W. (1994). « A spoken language for information retrieval ». In *Proceedings of ICSLP* : 1271-1274.

BOUSQUET-VERNHETTES C. (2002). *Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique*. Thèse de doctorat, Université de Toulouse III.

IDE N., VÉRONIS J. (1998). « Introduction to the Special Issue on WSD : The State of the Art ». In *Computational Linguistics* 24 (1) : 1-40.

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

ZOUAGHI A., ZRIGUI M., BEN AHMED M. (2005). « Un étiqueteur sémantique des énoncés en langue arabe ». In *Actes de RECITAL 2005* : 727-732.