L'influence du contexte sur la compréhension de la parole arabe spontanée

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

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

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

Résumé

Notre travail s'intègre dans le cadre du projet intitulé « Oréodule » : un système de reconnaissance, de traduction et de synthèse de la langue arabe. L'objectif de cet article est d'essayer d'améliorer le modèle probabiliste sur lequel est basé notre décodeur sémantique de la parole arabe spontanée. Pour atteindre cet objectif, nous avons décidé de tester l'influence de l'utilisation du contexte pertinent, et de l'intégration de différents types de données contextuelles sur la performance du décodeur sémantique employé. Les résultats sont satisfaisants.

Mots-clefs : analyse sémantique, modèle probabiliste, extraction automatique, contexte pertinent, information mutuelle moyenne.

Abstract

This work is part of a larger research project entitled « Oréodule » aiming to develop tools for automatic speech recognition, translation, and synthesis for the Arabic language. The core of our interest in this work is in improving the probabilistic model on which our semantic decoder rests. To achieve this goal, we tested the influence of the pertinent context use, and of the contextual data integration of different types, on the effectiveness of the semantic decoder. The results are satisfactory.

Keywords: semantic analysis, probabilistic model, automatic extraction, pertinent context, overage mutual information.

1. Introduction

Nos travaux s'inscrivent dans le cadre de la compréhension automatique de la langue arabe, et dans le contexte des communications homme-machine finalisées. L'utilisation des modèles statistiques pour la reconnaissance et la compréhension de la parole ont l'avantage de réduire fortement le recours à l'expertise humaine. En plus ils sont plus portables vers d'autres domaines ou vers des applications multilingues (Minker, 1999). Se baser sur de tels modèles pour l'interprétation sémantique des mots de l'énoncé, nécessite généralement la prise en considération du contexte d'énonciation. Les informations contextuelles jouent un rôle primordial dans la détermination du sens d'un mot dans un énoncé donné. Ces informations permettent de lever les ambiguïtés d'interprétation et améliorer les performances du système de compréhension (Bousquet-Vernhettes, 2002). Dans l'approche standard, le décodage du sens d'un mot est généralement déduit en analysant le contexte qui le précède et/ou qui le suit immédiatement. Or, dans le cas de la compréhension de l'arabe spontané, ceci n'est pas toujours optimum. Nous avons constaté par exemple, un taux d'erreur de 57 % en considérant un contexte de taille 1 (c'est-à-dire seulement le sens du mot qui précède le mot à interpréter sémantiquement). Afin de remédier à ce problème, nous avons décidé de ne considérer que les

sens des mots pertinents pour la sélection du Tse (notation d'ensemble de traits sémantiques) adéquat à la description de la signification du mot à interpréter. Nous tenons aussi compte du type de l'acte illocutoire accompli par l'énoncé auquel appartient le mot à interpréter (refus, demande, etc.) et de sa nature (par exemple demande de réservation, d'horaires, etc.) pour la prédiction du Tse à utiliser.

2. Les difficultés du décodage sémantique de la parole arabe

2.1. Diverses variétés de la parole arabe

La langue arabe est la sixième langue la plus parlée dans le monde avec environ 250 millions locuteurs. Pour des raisons historiques et idéologiques, cette langue présente une hiérarchisation de diverses variétés :

- L'arabe classique ou coranique : c'est la langue la plus ancienne, à partir de laquelle dérivent toutes les autres variantes (l'arabe dialectal et intermédiaire). L'étude de sa grammaire a commencé vers le milieu du XI^{e} siècle de l'hégire. C'est la langue écrite de la littérature et de la presse, et parlée ordinairement à la radio, dans les conférences et les discours officiels dans tous les pays arabes. L'arabe classique est appris à l'école.
- L'arabe intermédiaire : c'est une variante simplifiée de l'arabe classique et une forme élevée de l'arabe dialectal. Il emprunte son lexique aussi bien au dialecte qu'à l'arabe classique. Cette variante est aujourd'hui en plein essor, elle s'utilise de plus en plus couramment dans les enseignements et dans les médias. Elle permet de s'approcher des analphabètes et de la langue maternelle du peuple.
- L'arabe dialectal : c'est une autre variante de l'arabe classique. Il est essentiellement oral, c'est la langue de conversation quotidienne. Chaque pays arabe a son propre dialecte. Malgré qu'il existe plusieurs dialectes, l'intercompréhension est possible entre les différents pays.

Ces différents registres rendent le traitement automatique de la langue arabe, sous ses différentes variétés, impossible. C'est pour cela que les systèmes développés sont conçus pour traiter seulement l'une de ces variantes. Le décodeur sémantique que nous proposons dans cet article est dédié à la langue arabe classique.

2.2. Spécificités de la langue arabe

Le décodage sémantique de la parole arabe est une tâche très difficile vu sa richesse sémantique. Cette complexité est due aux spécificités de cette langue, qui sont :

- La liaison sans espace de la conjonction de coordination و (et) aux mots. Ceci rend difficile la distinction entre le و en tant qu'une lettre d'un mot (par exemple ورق « feuilles ») et le و ayant le rôle d'une conjonction de coordination (énoncé E). Or ce type de conjonction joue un rôle important dans l'interprétation d'un énoncé, en permettant d'identifier ses propositions.

  أريد معرفة توقيت القطار الذاهب إلى تونس وحجز مكان « ouridou ma'rifata tawqyta elqitar athaheb ila tuwnis wa hajza makan » → Je veux connaître l'horaire du train allant à Tunis et réserver une place. (E)

- L'ordre selon lequel est agencé les mots dans une phrase est variable : ceci complique la tâche de construction d'un modèle de langage approprié, à partir duquel sera interprété l'énoncé.
- La non voyellation de la plupart des textes arabes rencontrés dans les livres et les journaux rend la tâche d'apprentissage dans le cas d'utilisation d'un modèle probabiliste plus compliquée. En effet, au niveau sémantique la détection automatique du sens d'un mot non voyellé est très ambiguë. Il est même impossible de déduire le sens de certains mots non voyellés, si on ne connaît pas le contexte de leurs énonciations. Par exemple, le mot "مدرسة" (mdrst) peut avoir trois interprétations possibles selon la manière de sa voyellation. Il peut avoir la signification d'école, ou d'enseignante ou d'enseignée.

Ce qui complique davantage le traitement automatique de la langue arabe, est l'absence de formalisme théorique consistent capable de tenir compte de tous les phénomènes rencontrés lors de l'analyse sémantique de cette langue.

3. L'approche utilisée

3.1. Les méthodes couramment utilisées

Dans la littérature plusieurs méthodes ont été proposées pour le décodage sémantique de la parole spontanée. Certains utilisent les modèles de Markov cachés (Bousquet, 2002), d'autres les réseaux de neurones (Jamousi *et al.*, 2004), les modèles de langage N-grammes (Knight *et al.*, 2001), le λ-calcul ou encore les logiques, etc. Le tableau de la figure 1 ci-dessous, présente quelques formalismes utilisés pour la compréhension de la langue arabe et latine, ainsi que leurs principaux avantages et inconvénients.

Contrairement au latin, la compréhension automatique de la parole arabe spontanée reste encore très peu abordée au niveau de la recherche scientifique. Durant les deux dernières décennies les efforts ont été plutôt concentrés sur la réalisation des analyseurs morphologiques et syntaxiques pour l'arabe tel que (Ouersighni, 2001). Malgré l'importance de la représentation et de l'analyse sémantique pour la réalisation de n'importe quel système de compréhension, il n'existe que quelques travaux qui s'intéressent à ce domaine en vue du traitement automatique de la langue arabe écrite et non pas parlée (Haddad *et al.*, 2005 ; Meftouh *et al.*, 2001 ; Al-Johar *et al.*, 1997 ; Mankaï, 1996). Le système Al Biruni de (Mankaï, 1996) par exemple repose sur une combinaison du formalisme de la grammaire de cas de Fillmore et de la théorie sens texte de Mel'cuk pour l'analyse sémantique, la représentation du texte arabe et la manipulation de sa représentation. Quant à (Haddad *et al.*, 2005), il utilise la grammaire d'unification HPSG qui permet d'intégrer des connaissances syntaxique et sémantique dans une même grammaire, afin d'aboutir à une analyse profonde. Tous ces travaux cités s'intéressent plutôt au traitement de l'arabe écrit qu'à l'arabe parlé. La méthode que nous proposons dans ce papier, est inspirée de la grammaire des cas. Elle permet de décoder le sens des mots de l'énoncé en se basant sur les données contextuelles pertinentes, c'est à dire en ne considérant que le contexte qui possède une influence sémantique sur le mot à interpréter. L'avantage de notre méthode est que le contexte est déterminé automatiquement et ne nécessite pas l'intervention d'un expert humain. En plus les données contextuelles qui contribuent à l'interprétation d'un mot sont de plusieurs types : illocutoires, linguistiques et sémantiques. La considération de divers types de données contextuelles nous a permis d'améliorer la performance de notre décodeur. En plus, notre modèle est adapté au traitement de l'arabe parlé, puisqu'il ne repose que sur l'analyse des éléments porteurs de sens présents dans la requête du locuteur. Les éléments redondants ou

non significatifs (les termes d'appui du discours tels que إذا (alors) ou نعم (oui), ou le tic « hein » utilisé souvent à la suite d'une question ou d'une suggestion par la plupart des locuteurs, etc.) sont ignorés lors de l'analyse. Ils sont éliminés lors de la phase du prétraitement de l'énoncé reconnu. Dans les paragraphes suivants nous présentons l'approche que nous proposons pour le décodage sémantique de la parole arabe spontanée, et la manière d'extraction automatique du contexte pertinent.

Formalismes utilisés	Systèmes ou projets utilisant ce formalisme	Plus adapté au traitement de	Avantages	Inconvénients
HMM (Hidden Markov Models)	(Bousquet, 2002) (Minker, 1999)	L'oral	Existence d'algorithmes puissants (tels que Viterbi, A*) permettant de déterminer la solution optimale.	Nécessite de corpus de tailles assez volumineuses.
Réseaux de neurones (Neuronal Networks)	(Jamousi *et al.*, 2004), (Meftouh *et al.*, 2001)	L'oral et l'écrit	Capacité de généralisation et de flexibilité.	Développement coûteux en temps et structures générés très complexes. Nécessite comme les HMM de corpus de tailles assez volumineuses.
HPSG (Head Phrase Structure Grammar	(Haddad *et al.*, 2005) le projet Verbmobil	L'écrit (car elle permet d'analyser une phrase en terme de constituants syntaxiques)	Permet une intégration plus explicite dans une seule structure des différents niveaux de l'analyse linguistique : phonétique, syntaxique, et sémantique.	N'est pas adaptée pour être utilisée dans un système vocal interactif.
Grammaire de cas	(Mankaï, 1996) (Minker, 1999)	L'oral	Autorise le traitement des phrases agrammaticales et nécessite moins d'expertise en linguistique.	Réduit le rôle de la syntaxe.

Figure 1. Exemple de formalismes utilisés pour la compréhension du langage naturel

3.2. Les caractéristiques de notre approche

Pour réaliser notre décodeur sémantique, nous avons opté pour les choix suivants :

- Une représentation componentielle du sens des mots : chaque mot significatif pour l'application est représenté à partir d'un ensemble de traits sémantiques noté Tse = {domaine, classe sémantique, trait micro sémantique} et un ensemble de traits syntaxiques noté Tsy = {genre, nombre, nature}. Les traits de l'ensemble Tse indiquent respectivement le domaine de l'application, la classe sémantique à laquelle appartient le mot à interpréter, et le dernier trait c'est un trait micro sémantique qui permet de différencier le sens des mots appartenant à une même classe sémantique. Nous signalons que les mots polysémiques ou possédant un même rôle sémantique possèdent le même ensemble de traits Tse. En appliquant cette représentation, le sens du mot الذاهب ″allant″ par exemple est décrit comme suit : الذاهب « ethaheb" → Tse = {(transport) نقل ″naql″, (mouvement) حركة ″haraka″, (destination) وجهة ″wijha″} + Tsy = {(masculin) مذكر, (singulier) مفرد, (nom) اسم}.
- Une analyse sélective : pour le décodage sémantique des énoncés, nous nous sommes basés plutôt sur une analyse sémantique et nous avons considéré seulement les éléments significatifs pour l'application. Les mots vides sont éliminés lors de la phase du prétraitement de l'énoncé. Cette analyse est plus tolérante aux erreurs grammaticales

qui caractérisent la parole spontanée. En plus, elle ne nécessite pas des connaissances linguistiques très approfondies.

- Une méthode anthropocentrée basée sur une analyse de corpus : pour la construction de notre structure de représentation de sens SRS (Zouaghi *et al.*, 2004), nous avons développé une méthode basée sur une analyse de corpus pour l'extraction des mots significatifs, des mots de référence et des classes sémantiques de l'application, et sur une coopération homme/machine pour l'interprétation des mots. Les mots vides sont éliminés en utilisant un filtre lexical. Selon cette méthode le rôle de l'utilisateur est de définir et d'attribuer l'ensemble des Tse et de Tsy aux mots. Et le rôle de la machine est de satisfaire les contraintes d'intégrités afin d'aboutir à une SRS non ambiguë et cohérente. Notre système se base en tout sur une dizaine de contraintes. Un exemple de contrainte à vérifier est que : deux mots différents ne peuvent pas être décrits par un même ensemble de Tse sauf dans le cas où ils sont considérés comme synonymiques ou possédant un même rôle sémantique. Cette méthode nous a permis de faciliter la tâche d'interprétation des mots, ainsi que de la tâche de construction de la SRS et de la maintenance de sa cohérence. La figure 2 suivante, présente l'interface homme/machine utilisée pour interpréter chaque mot via Tse et Tsy.

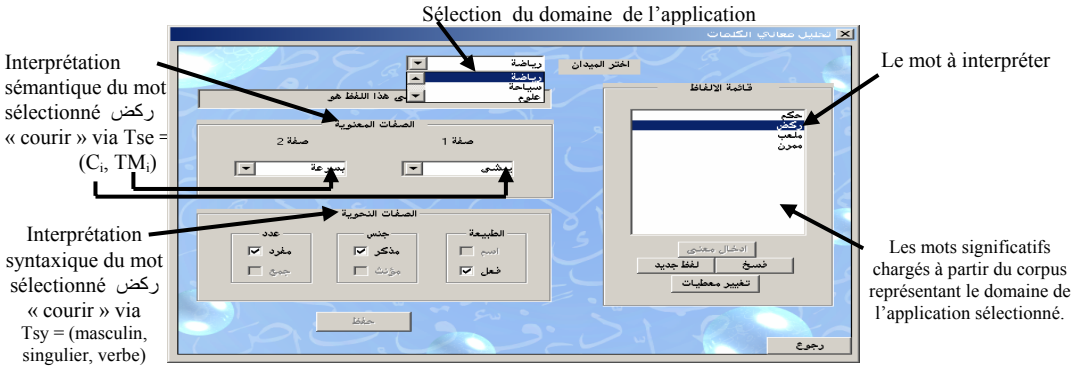

Figure 2. L'interface homme/machine d'interprétation des mots significatifs d'une application

- Une grammaire probabiliste : cette grammaire participe au choix des Tse adéquats à la description des mots constituant l'énoncé à interpréter. Cette grammaire permet de tenir compte de plusieurs informations contextuelles en même temps. En plus elle ne considère que les Tse pertinents déjà utilisés pour la prédiction du Tse correspondant à un mot pas encore interprété. Notre modèle permet de contraindre l'analyse sémantique des énoncés reconnus, en réduisant l'espace de recherche du décodage sémantique des énoncés. Ceci est réalisé en se reposant sur une estimation des probabilités d'interprétation d'un mot donné, sur les mots qui agissent sur son sens (en utilisant la notion d'information mutuelle moyenne), et sur l'utilisation de modèles de type POS tagging pour la détermination de chacun des traits de l'ensemble Tse. L'utilisation pour l'interprétation d'un mot donné, les mots qui agissent sémantiquement sur ce dernier, permet de surmonter les problèmes de l'oral spontané. Ceci a été prouvé à travers le formalisme des grammaires de cas de Fillmore. Quant aux modèles POS tagging, leur performance a été démontrée dans le domaine de l'analyse syntaxique. Ci-dessous l'équation exprimant la probabilité d'interprétation d'un mot Mi par le couple (Ci, TMi) en tenant compte du type de l'énoncé. On remarque que dans cette formule nous n'avons pas considéré le domaine de l'application puisqu'il est prédéfini à l'avance. Dans notre cas, il s'agit du domaine des renseignements ferroviaires. Les

approximations et les assomptions que nous avons considérées pour l'obtention de ce modèle (décrit par la formule 1) sont détaillées dans (Zouaghi *et al.*, 2005a).

$$P((C_i, TM_i)/ M_i, NT_j) = P(NT_j/ Mr) \times P(C_i / NT_j, M_{i-1}, CP_{i-1}, CP_{i-2}) \times P(TM_i / Ci, TseP_{i-1}) \quad (1)$$

On remarque bien que cette probabilité est calculée en fonction du produit de trois probabilités conditionnelles. La première probabilité *P (NTj/ Mrk)* permet d'identifier le type de l'énoncé, s'il s'agit d'une demande de réservation, d'annulation de billet, etc. Ceci en tenant compte des mots de références Mrk présents dans l'énoncé du locuteur. Les mots de références sont des uni-grammes, des bi-grammes ou des tri-grammes (qui peuvent être distants) dont les probabilités d'occurrence sont égales à un. Par exemple le bi-gramme أريد حجز qui correspond au tri-gramme « je veux réserver » en français est un mot de référence indiquant qu'il s'agit d'une demande de réservation. La deuxième probabilité P(Ci / NTj, Mi-1, CPi-1, CPi-2) permet de déterminer la classe sémantique Ci à laquelle appartient le mot à interpréter Mi, en tenant compte du type de l'énoncé et des deux classes sémantiques pertinentes précédentes. Et la troisième probabilité P(TMi / Ci,TsPi-1) permet de déterminer le trait micro sémantique TMi à attribuer à Mi, en tenant compte de la classe qui a été attribuée à ce mot et du Tse pertinent précédent (voir paragraphe 4 en ce qui concerne la méthode d'extraction des Tse pertinents).

3.3. Principe du décodage sémantique

Nous entendons par décodage sémantique d'un énoncé, l'étiquetage de chacun de ses mots significatifs via un ensemble Tse. Comme le montre la figure 3, le décodage sémantique de l'énoncé prétraité repose sur un modèle de langage probabiliste (Zouaghi *et al.*, 2005b) et un lexique sémantique. Le modèle probabiliste contribue à la sélection des Tse à affecter aux mots de l'énoncé à interpréter, et le lexique sémantique décrit le sens de chaque mot via un ensemble de Tse et un ensemble de Tsy. À partir de l'énoncé décodé est déduit son sens. Ceci en remplissant les attributs du schéma identifié avec les valeurs correspondantes.

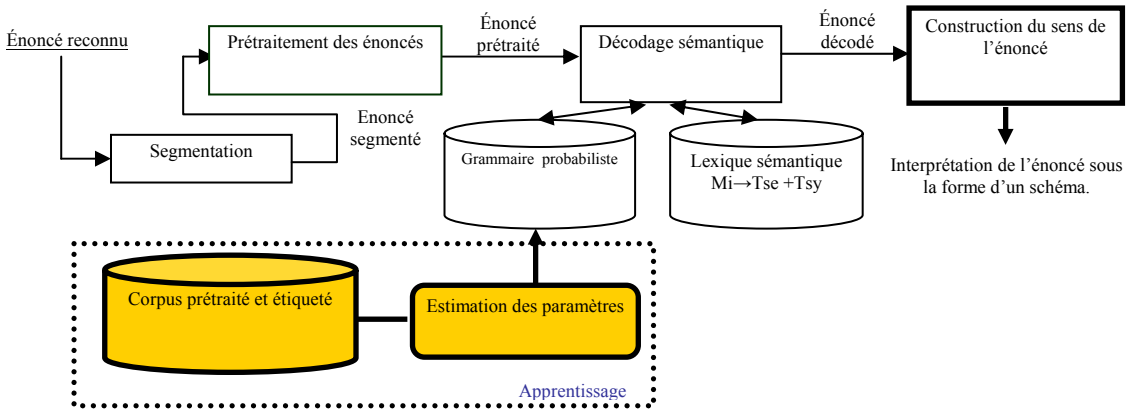

Figure 3. Architecture de l'analyseur sémantique

Au cours de la phase d'apprentissage nous avons considéré un corpus étiqueté et prétraité pour l'estimation des paramètres du modèle probabiliste. Le prétraitement du corpus représentant l'application nous a permis de simplifier la complexité et réduire la taille du modèle probabiliste. Ce prétraitement consiste comme pour le prétraitement de l'énoncé à éliminer par exemple les mots vides, à regrouper certains mots en une seule entrée, etc.

4. Extraction du contexte pertinent pour le décodage sémantique

4.1. Principe de l'extraction

Pour déterminer l'ensemble de Tse à affecter aux mots de l'énoncé à décoder sémantiquement, nous utilisons dans notre modèle probabiliste seulement les ensembles de Tse pertinents. Nous entendons par Tse pertinents, les Tse utilisés pour la description du sens des mots possédant une affinité sémantique forte avec le mot Mi. Ainsi pour identifier la classe sémantique Ci à laquelle appartient le mot Mi, nous considérons dans l'équation (1) les deux classes sémantiques CP_{i-1} et CP_{i-2} des deux Tse affectés aux deux mots ayant les plus grandes affinités sémantiques avec Mi. Pareil pour déterminer le trait micro sémantique TMi permettant de différencier le sens de Mi avec les autres mots appartenant à la même classe Ci que Mi, nous considérons dans (1) seulement l'ensemble TsePi = {CPi, TMPi} qui a été affecté au mot ayant la plus grande affinité sémantique avec Mi. Pour atteindre cet objectif, nous nous sommes basés sur la notion d'information mutuelle moyenne (Rosenfeld, 1994) qui permet de calculer le degré de corrélation ou de co-occurrence de deux mots donnés.

4.2. Calcul de l'affinité sémantique

Considérons un énoncé E à interpréter : E = M1 M2 …Mi… Mn. Soit ME = {ME_{-K}, .., ME_{-1}, ME_1, .., ME_k} : l'ensemble de mots entourant le mot Mi à interpréter en considérant une fenêtre de taille k. Comme notre modèle considère que le contexte droit de Mi pour le choix du Tse à affecter à ce mot. L'ensemble ME est ainsi réduit à MEd = {M_1 M_2 …M_{i-1}}, qui est l'ensemble de mots précédant Mi dans l'énoncé. Pour déterminer maintenant l'affinité sémantique la plus forte entre Mi et son contexte, nous commençons par calculer les informations mutuelles moyennes entre Mi et chacun des mots appartenant à MEd. La formule de l'information mutuelle moyenne IMm (Rosenfeld, 1994) est la suivante :

$$IMm(Mi, MEdj) = P(Mi, MEdj)\times \text{Log}\,[P(Mi / MEdj) / P(Mi).P(MEdj)] + P(\overline{Mi}, \overline{MEdj})\times \text{Log}\,[P(\overline{Mi} / \overline{MEdj}) / P(\overline{Mi}).P(\overline{MEdj})] + P(\overline{Mi}, MEdj)\times \text{Log}\,[P(\overline{Mi} / MEdj) / P(\overline{Mi}).P(MEdj)] + P(Mi, \overline{MEdj})\times \text{Log}\,[P(Mi / \overline{MEdj}) / P(Mi).P(\overline{MEdj})] \text{ ; avec } 1\leq j \leq i-1 \quad (2)$$

Nous avons préféré utiliser l'IMm (équation 2) plutôt que l'information mutuelle classique, car l'IMm permet de calculer en plus l'impact de l'absence d'un mot sur l'apparition de l'autre. L'affinité sémantique la plus forte ou maximale *AffM* que possède le mot Mi avec son contexte droit est déterminée alors à partir de la formule (3) suivante :

$$AffM(Mi, MEd) = \max_{1\leq j\leq i-1} IMm(Mi, MEDj) \quad (3)$$

5. Application du modèle et résultats

Nous avons utilisé une centaine d'énoncés spontanés (obtenus par la méthode du magicien d'Oz) différents de ceux du corpus d'apprentissage. Le corpus d'apprentissage (constitué de 10 000 énoncés représentant le domaine des renseignements ferroviaires) a été étiqueté avec 37 ensembles Tse différents. Pour juger de la qualité de notre décodeur, nous avons calculé le pourcentage des Tse qui sont incorrectement attribuées, à partir de la formule suivante : $Terreur=N_{inc}/N\times 100$. Où, N_{inc} est le nombre de Ts incorrectement attribués, et N est le nombre total des Tse attribués par un expert au corpus de test. N est égal à 500 dans ce test.

La figure 4 ci-dessous montre les Terreur obtenus en considérant des modèles bi-classes et tri-classes classiques dans un premier temps et notre modèle hybride défini dans un deuxième temps. Les modèles bi-classes et tri-classes sont des modèles de type POS tagging. Le modèle bi-classes consiste en la considération de la classe C_{i-1} affectée au mot M_{i-1} pour la détermination de la classe C_i à affecter à M_i qui succède directement le mot M_{i-1}. Et le modèle tri-classes consiste en la considération des classes C_{i-2} et C_{i-1} affectées successivement aux mots M_{i-2} et M_{i-1} pour la détermination de C_i à affecter à M_i qui succède directement les mots M_{i-2} et M_{i-1}. La longueur de l'historique est fixée à 3 pour la détermination des C_i et à 2 pour TM_i. Pour chaque type de modèle, nous avons calculé l'influence des informations lexicales sur la performance du décodeur. Pour notre modèle hybride, nous avons calculé en plus l'influence de la considération des Tse pertinents.

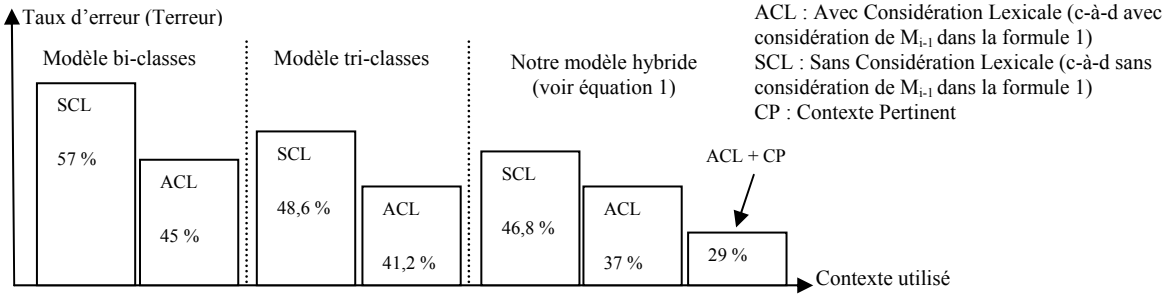

Figure 4. Taux d'erreurs obtenus en considérant de différentes données contextuelles

D'après le tableau ci-dessus, nous remarquons que chaque fois que l'on considère 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 intégrant en plus le type de l'énoncé, le taux d'erreur atteint 37 % (avec considération des données lexicales). En considérant ensuite les Tse pertinents pour la prédiction du Tse décrivant le sens du mot à interpréter, nous remarquons que nous atteignons un taux d'erreur de l'ordre de 29 %. En analysant notre corpus de test, nous avons remarqué que ce taux d'erreur est du principalement aux énoncés ayant 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 plus ouvertes.

6. Conclusion

Nous avons présenté dans cet article un décodeur sémantique basé sur un modèle de langage hybride, qui permet d'intégrer des données contextuelles lexicales, sémantiques ainsi qu'illocutoire en même temps. Il permet en plus de ne tenir compte que des ensembles de traits sémantiques Tse pertinents dans l'historique du mot à interpréter. Pour cela, nous avons développé une méthode permettant d'extraire automatiquement ces Tse pertinents qui décrivent les sens des mots ayant une influence sémantique sur le mot à interpréter. Ceci est atteint, en se basant sur la notion d'information mutuelle moyenne de Rosenfeld. Les résultats trouvés sont satisfaisants. Dans le prochain avenir nous comptons évaluer notre modèle par rapport aux modèles dits distants ou les modèles obtenus par combinaison linéaire de modèles de langage bien connus comme le maximum d'entropie.

Références

AL-JOHAR, B., MCGREGOR, J. (1997). « A Logical Meaning Representation for Arabic (LMRA) ». In *Actes du 15th National Computer Conference* : 31-40.

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 ». In *Actes de TALN*.

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

HADDAD B., YASEEN M. (2005). « A Compositional Approach Towards Semantic Representation and Construction of ARABIC ». In *Actes de LACL*.

JAMOUSSI S. (2004), *Méthodes statistiques pour la compréhension automatique de la parole*. Thèse de doctorat de l'Université Henri Poincaré.

KATZ S.M. (1987). « Estimation of probabilities from sparse data for the language model component of a speech recognizer ». In *IEEE Transactions on Acoustics* : 400-401.

KNIGHT S., GORELL G., RAYNER M., MILWARD D., KOELING R., LEWIN I. (2001). « Comparing grammar-based and robust approaches to speech understanding : a case study ». In *Actes de European conference on speech communication and technology* : 1779-1782.

MANKAÏ Naanaa C. (1996). *Compréhension automatique de la langue arabe. Application : Le système Al Biruni*. Thèse de doctorat de l'Université de Tunis II .

MEFTOUH K., LASKRI M.T. (2001). « Generation of the Sense of a Sentence in Arabic Language with a Connectionist Approach ». In *Actes de AICCSA*.

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

OUERSIGHNI R. (2001), « A major offshoot of the Dinar-MBC project : AraParse, a morphosyntactic analyzer for unvowelled Arabic texts ». In *Actes de ACL/EACL*.

PEPELNJAK K., GROS J., MIHELIC F., PAVEŠIC N. (1995). « Linguistic Analysis in a Slovenian information retrieval system for flight services ». In *Actes du Workshop on Spoken Dialogue Systems* : 65-68.

ROSENFELD R. (1994). *Adaptive statistical language modelling : A maximum entropy approach*. Thèse de doctorat de l'Université de Carnegie Mellon.

SENEFF S. (1992). « Robust parsing for spoken language systems ». In *Actes de ICASSP* : 189-192.

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

ZOUAGHI A., ZRIGUI M., BEN AHMED M. (2005b). « A statistical model for semantic decoding of Arabic language statements ». In *Actes de NODALIDA* : 93-95.