Analyse sémantique des énoncés oraux arabes dans un contexte de dialogue homme-machine

Younès Bahou, Houssem Safi, Lamia Hadrich Belguith

bahou.younes@caramail.com, safi.houssem@isimsf.rnu.tn, l.belguith@fsegs.rnu.tn.
Laboratoire LARIS-MIRACL, FSEGS – Université de Sfax, Tunisie.

ABSTRACT

In this paper we present the ASAR system of semantic analysis of Arabic speech statements in a context of human-machine oral dialogue.. ASAR is based on a parsing method that is strongly guided by semantics and that uses the case grammar formalism. This method consists of three main steps: a step of pretreatment and semantic tagging, a step of semantic pattern identification and filtering and a step of pattern generation.

Keywords: semantic analysis, spontaneous Arabic speech processing, human-machine dialogue, case grammars.

1. INTRODUCTION

Actuellement, il existe plusieurs Systèmes de Dialogue Oral Homme-Machine (SDOHM) opérationnels. Nous citons à titre d'exemple le système Jupiter pour des renseignements météorologiques en anglais [11] et le système Ritel permettant à un utilisateur de rechercher en français oral des informations générales [7]. Pour le cas de l'arabe, à notre connaissance, il n'existe aucun serveur vocal capable de dialoguer en arabe. Cela est principalement dû au manque d'outils d'aide à l'élaboration de tels systèmes. En effet, la plupart des outils existants restent à l'échelle de prototypes de laboratoire. Nous citons à titre d'exemple l'outil de Satori pour la reconnaissance des chiffres arabes basée sur CMUSphinx [5] et l'outil de Ramsay pour la synthèse de la parole arabe [4].

Par ailleurs, très peu de chercheurs se sont intéressés à l'analyse sémantique de l'arabe écrit ou parlé. Dans cet article, nous présentons notre méthode d'analyse sémantique que nous appliquons dans le cadre du système ASAR d'Analyse Sémantique des énoncés oraux ARabes dans un contexte de dialogue oral homme-machine. Comme domaine d'application, nous avons choisi celui des renseignements sur le transport ferroviaire tunisien.

2. DIFFICULTES DE L'ANALYSE SEMANTIQUE DE L'ORAL ARABE

L'objectif d'un analyseur sémantique de la parole est la construction de la représentation sémantique de l'énoncé prononcé par l'usager. Cette représentation est calculée en fonction de la séquence des mots reconnus, du cotexte et du contexte. Une analyse sémantique fiable et robuste doit faire face aux nombreux problèmes dus à la nature incertaine des énoncés oraux, aux erreurs produites par le module de reconnaissance de la parole, à la nature spontanée de l'interaction, etc. Les caractéristiques intrinsèques au dialogue oral qui influent sur l'analyse sémantique de la parole sont principalement : la spontanéité des énoncés et la structure agrammaticale des énoncés. En plus de ces caractéristiques, la langue arabe parlée présente une hiérarchisation de plusieurs variétés à savoir, i) l'arabe moderne qui constitue la langue écrite du Coran. Il s'agit de la variété retenue comme langue officielle dans tous les pays arabes. ii) les arabes dialectaux parlés dans les pays arabes. Notons que ces dialectes varient d'un pays à un autre, voire même d'une région à une autre. La différenciation porte sur la prosodie, la phonologie, la morphologie et la syntaxe. Dans le cadre du présent travail, nous nous intéressons à l'arabe moderne parce qu'il est difficile pour un analyseur sémantique de traiter différents dialectes et aussi parce qu'il y a quasi-absence d'outils pour les dialectes arabes.

3. APPROCHES D'ANALYSE SEMANTIQUE POUR LES SDOHM

Différentes approches d'analyse sémantique pour les SDOHM ont été proposées, dans la littérature, et testées dans divers contextes applicatifs.

3.1 Approche guidée par la syntaxe

Dans cette approche, une analyse syntaxique robuste et partielle est envisagée. En effet, en analyse automatique de la parole, il est apparu qu'une analyse syntaxique complète, pouvait faire décroître les performances en présence des mots inconnus, des erreurs de reconnaissance et des phénomènes dus à la spontanéité de l'interaction.

Plusieurs systèmes se sont basés sur cette approche. Nous citons à titre d'exemples le système ROMUS [3] et le système LOGUS [8]. Il s'agit de deux systèmes d'analyse sémantique de la parole appliquée au domaine de renseignement touristique. Ces deux systèmes reposent sur une analyse partielle par segments.

3.2 Approche guidée par la sémantique

Dans cette approche, l'analyse sémantique est limitée à la recherche du sens dit utile de l'énoncé. Elle repose sur l'identification de séquences-clés à partir desquels une structure sémantique prédéfinie sera instanciée. Donc, l'idée consiste à analyser uniquement les parties jugées pertinentes de l'énoncé. Plusieurs systèmes reposent sur cette approche nous citons, à titre d'exemple, le système PHOENIX [9] qui offre des renseignements sur le transport aérien. Il comporte un analyseur sémantique flexible fondé sur la grammaire des cas et compilé dans un ensemble de réseaux de transitions récursifs. Pour l'analyse sémantique de la parole arabe, nous citons le décodeur sémantique de Zouaghi [10] appliqué au domaine de renseignement ferroviaire. Ce décodeur utilise une grammaire probabiliste qui permet de tenir compte de plusieurs informations contextuelles.

Cette approche s'avère efficace pour des domaines très limités où l'ambiguïté sémantique est réduite. En plus, elle présente une certaine robustesse face aux difficultés causées par la spontanéité des énoncés oraux, l'agrammaticalité du langage parlé et la présence de mots non reconnus.

3.3 Approche mixte

Dans cette approche, une analyse syntaxique complète est tout d'abord envisagée. En cas d'échec, des techniques sémantiques sont alors utilisées. Le système TINA [6] basé sur cette approche offre un service de renseignement sur le transport aérien aux États-Unis. L'analyseur sémantique de TINA se base sur une grammaire hors-contexte contrainte. Il s'agit d'une analyse descendante basée sur des règles syntaxiques auxquelles s'ajoutent des contraintes sémantiques. La grammaire est transformée automatiquement en un automate probabilisé qui permet d'avantager les constructions les plus fréquentes. Devant l'insuffisance de cette analyse complète sur l'oral spontané, une stratégie d'analyse robuste par *chart* étendu a été intégrée. Cette analyse permet le recouvrement des groupes analysés correctement dans l'énoncé, en cas d'échec de l'analyse complète.

4. NOTRE MÉTHODE

Notre méthode d'analyse sémantique de la langue arabe est basée sur le formalisme de grammaire des cas [2] et permet l'extraction du sens utile contenu dans les mots ou dans une suite de mots de l'énoncé. Ce sens utile sera représenté sous forme de schémas sémantiques. Cette méthode est fortement guidée par la sémantique. Elle consiste en trois principales étapes : une étape qui permet le prétraitement et l'étiquetage sémantique de l'énoncé reconnu, une étape qui permet d'identifier et de filtrer les schémas et une étape qui permet de générer les schémas sélectionnés (voir figure 1).

4.1 Prétraitement et étiquetage sémantique

4.1.1 Prétraitement

Les problèmes spécifiques à la langue arabe que nous avons décrit dans la section 2 rendent la structure déclarative rigide et difficilement maîtrisable. Cette situation introduit également des ambiguïtés qui peuvent provoquer des erreurs d'analyse. C'est pourquoi, dans notre méthode, chaque énoncé subit un prétraitement qui vise à normaliser l'énoncé afin de faciliter les étapes ultérieures. Il consiste à faire une analyse morphologique embryonnaire, à traiter les autocorrections, à traiter les répétitions, etc.

Traitement des répétitions. Ce traitement consiste à éliminer les répétitions et ce par la suppression d'une occurrence du mot ou du groupe de mots qui se répète. Il est basé sur un ensemble de règles. A titre d'exemple prenons l'énoncé (1) qui présente une répétition du mot ' صفاقس ' (Sfax).

(1) ' أريد ثمن تذكرة ذهاب آه لا ذهاب وإيّاب من **صفاقس** **صفاقس** إلى تونس على الساعة الحادية عشر وثلاثة وثلاثون دقيقة '

(Je voudrais connaître le prix d'un billet aller simple eh non aller retour de **Sfax Sfax** vers Tunis à onze heures trente trois minutes)

L'énoncé (1) sera transformé en (2).

(2) ' أريد ثمن تذكرة ذهاب آه لا ذهاب وإيّاب من صفاقس إلى تونس على الساعة الحادية *عشر وثلاثة وثلاثون* دقيقة '

(Je voudrais connaître le prix d'un billet aller simple eh non aller retour de Sfax vers Tunis à onze heures trente trois minutes)

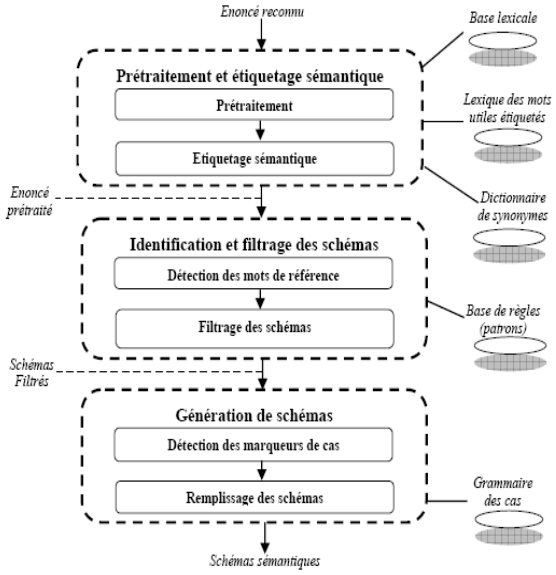

Figure 1 : Etapes de notre méthode

Traitement des autocorrections. Ce traitement consiste à comparer chaque mot avec le mot qui le suit. S'ils ont la même étiquette sémantique alors le premier mot est éliminé. De plus, si dans un énoncé, il existe un marqueur de rectification alors le segment

qui précède ce marqueur sera remplacé par celui qui le succède.

L'énoncé (2) présente une correction du segment ' ذهاب ' (aller simple) erroné par le segment ' ذهاب وإيّاب ' (aller retour). Cette correction est introduite par le marqueur de rectification ' آه لا ' (eh non) qui contient une information sémantique importante. L'énoncé suivant est le résultat du traitement de l'autocorrection de l'énoncé (2).

(3) ' أريد ثمن تذكرة ذهاب وإيّاب من صفاقس إلى تونس على الساعة الحادية عشر وثلاثة وثلاثون دقيقة '

(Je voudrais connaître le prix d'un billet aller retour de Sfax vers Tunis à onze heures trente trois minutes)

Traitement des nombres représentant un horaire. Ce traitement consiste à détecter dans un énoncé un marqueur d'horaire (e.g., ' نحو+الساعة '(vers+l'heure)) et un contre marqueur d'horaire (e.g., ' دقيقة ' (minute)) s'il existe. Ensuite, ce traitement procède à la transformation des mots, qui désignent l'horaire et qui se trouvent entre ces deux marqueurs en un nombre numérique. Ainsi, L'énonce (3) sera transformé en l'énoncé (4).

(4) ' أريد ثمن تذكرة ذهاب وإيّاب من صفاقس إلى تونس على الساعة **11** و **33** دقيقة '

(Je voudrais connaître le prix d'un billet aller retour de Sfax vers Tunis à **11** heures **33** minutes)

4.1.2 Étiquetage sémantique

Cette phase consiste à donner une étiquette sémantique pour chaque mot, porteur de sens ou marqueur de cas, en se basant sur un lexique des mots utiles étiquetés. A titre d'exemple l'énoncé (4) sera étiqueté sémantiquement comme suit :

(5)

<*Type_Billet* , ذهاب وإيّاب> <*Billet* , تذكرة> <*Marq_Tarif* , ثمن>
<*Ville_Départ* , صفاقس> <*Marq_Ville_Départ* , من>
<*Ville-Arrivée* , تونس> <*Marq_Ville_Arrivée* , الى>
<*Minute* , 33> <*Heure* , 11> <*Marq_Heure* , على الساعة>
<*Marq_Minute* , دقيقة>

4.2 Identification et filtrage des schémas

L'identification des schémas sémantiques se base sur la détection des mots de référence. Le filtrage des schémas se base sur l'occurrence des mots de référence dans ces derniers.

La détection des mots de référence. A chaque schéma correspond une liste exhaustive de mots de référence. Ces mots permettent l'identification des schémas sémantiques correspondant à la requête de l'interlocuteur. En effet, les mots de référence d'un schéma peuvent être des mots simples ou des mots composés de deux mots de référence.

Le mot ' ثمن ' (prix) et la combinaison des deux mots ' ثمن+تذكرة ' (prix+billet) de l'énoncé (5) font référence au schéma *Tarif_Voyage* de notre grammaire. Aussi, le mot ' تذكرة ' (billet) fait référence au schéma *Réservation_Billet*.

Filtrage des schémas. Un énoncé peut contenir plusieurs mots de référence qui peuvent appartenir à des schémas différents. En effet, cette phase d'analyse permet de sélectionner le(s) schéma(s) correspondant à l'énoncé traité. Ainsi, le filtrage se fait en parcourant tous les schémas sémantiques et en calculant pour chacun d'eux un score d'apparition des mots de référence. Ce score correspond au nombre de mots de référence d'un schéma donné dans l'énoncé. Ainsi, le schéma retenu est celui ayant le score le plus élevé. Si plusieurs schémas ont le même score, ils seront tous retenus. Pour l'énoncé (5) le schéma retenu est le schéma *Tarif_Voyage* puisqu'il a obtenu le score le plus élevé (i.e., le score du schéma *Tarif_Voyage* est égale à 2 alors que le score du schéma *Réservation_Billet* est égale à 1).

4.3 Génération de schémas

L'étape de génération de schémas sémantiques consiste d'abord à détecter les marqueurs de cas ensuite à remplir les schémas.

La détection des marqueurs de cas. Chaque schéma est composé de cas et de sous cas. Un cas est identifié grâce aux marqueurs de cas. En effet, nous distinguons deux types de marqueurs de cas : les pré-marqueurs et les post-marqueurs. Dans l'énoncé (5), le mot ' ساعة ' (heure) est un pré-marqueur de l'heure et le mot ' دقيقة ' (minute) est un post-marqueur des minutes. Ces deux marqueurs permettent l'identification des deux informations relatives à l'heure et aux minutes dans l'énoncé (5).

Remplissage des schémas. Le remplissage de ces schémas se base sur les marqueurs des cas identifiés et sur un ensemble de règles spécifiques (dans le cas d'absence de marqueurs de cas). La figure 2 représente le schéma *Tarif_Voyage* généré après le traitement de l'énoncé (5).

<Schéma *Tarif_Voyage*>
  <Itinéraire>
    <Ville_Départ>صفاقس</Ville_Départ>
    <Ville_Arrivée>تونس</Ville_Arrivée>
  </Itinéraire>
  <Type_Train>$</Type_Train>
  <Classe_Train>$</Classe_Train>
  <Type_Billet>ذهاب وإيّاب</Type_Billet>
  <Jour_Voyage>$</Jour_Voyage>
  <Heure_Voyage>11:33</Heure_Voyage>
  <Date_Voyage>$</Date_Voyage>
</Schéma *Tarif_Voyage*>

Figure 2 : Schéma correspondant à l'énoncé (5)

5. LE SYSTÈME ASAR

Le système ASAR est un Analyseur Sémantique d'énoncés oraux ARabes pour le domaine de

renseignements sur le transport ferroviaire de la Société Nationale des Chemins de Fer Tunisiens (SNCFT). Il est basé sur le formalisme de grammaire des cas et repose sur la méthode d'analyse que nous avons proposée. La grammaire des cas de ASAR comporte six schémas : *Tarif_Voyage*, *Horaire_Voyage*, *Réservation_Billet*, *Durée_Voyage*, *Trajet_Train* et *Type_Train*. Chaque schéma contient des mots de référence et des cas sémantiques relatifs à notre domaine applicatif.

Pour réduire le nombre de mots de référence et de marqueurs de cas, ASAR utilise un dictionnaire de synonymes. Afin de limiter la taille de la base lexicale, chaque mot du lexique est réduit à sa forme canonique avant d'être ajouté à la base. Ainsi, ASAR utilise le système MORPH 2 [1] pour déterminer les formes canoniques des mots.

ASAR est programmé sous l'environnement *JBuilder 10* avec le langage *JAVA*. La base lexicale, le lexique des mots utiles étiquetés, le dictionnaire de synonymes, la base des règles et la grammaire des cas sont stockés dans des fichiers *XML*.

6. EVALUATION

Vu que les ressources linguistiques arabes sont très rares voire même indisponibles, nous étions amenés à construire notre propre corpus d'évaluation (1003 requêtes soit 12321 mots) selon la technique du *Magicien d'Oz*. Ainsi, nous avons utilisé des scénarios traitants des renseignements sur le transport ferroviaire tunisien.

L'évaluation de ASAR a montré que ce système génère 186 erreurs (soit en moyenne une erreur par 5 énoncés). 67 erreurs sont dues à un échec de détermination de la forme canonique par le système MORPH 2. Les mesures de rappel, de précision et de F-Measure sont respectivement de 73.00%, 70.62% et 71.79%. Notons que le temps moyen d'exécution d'un énoncé est égal à 0.279 secondes.

Le taux d'erreur de 18.54% s'explique principalement par la présence, dans les énoncés, des mots tronqués et des mots mal reconnus. Ces deux phénomènes ne sont pas encore pris en considération par notre système.

7. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons proposé une méthode d'analyse sémantique de la langue arabe dans un contexte de dialogue oral homme-machine. Nous avons aussi présenté le système ASAR qui se base sur la méthode proposée. Les résultats d'évaluation de ASAR sont très encourageants même si le système ASAR, à son état actuel, ne traite pas quelques problèmes tels que : les mots tronqués et les mots mal reconnus. En effet, nous avons obtenu 71.79% pour la mesure F-Measure.

Comme perspectives, nous envisageons d'évaluer ASAR sur un corpus plus grand. Aussi, nous projetons d'étudier et d'apporter des solutions aux phénomènes cités concernant les mots tronqués et les mots mal reconnus.

BIBLIOGRAPHIE

[1] H. L. Belguith and N. Chaaben. Analyse et désambigüisation morphologiques de textes arabes non voyellés. *TALN'06*, Leuven, Belgique, 2006.

[2] C. J. Fillmore. *The case for case*. Universals in Linguistic Theory, New York, 1968.

[3] J. Goulian, J.-Y. Antoine and F. Poirier. How NLP techniques can improve speech understanding : ROMUS – a Robust Chunk based Message Understanding System Using Link Grammars. *EUROSPEECH'03*, Geneva, 2003.

[4] A. Ramsay and H. Mansour. Towards including prosody in a text-to-speech system for modern standard Arabic. *Computer Speech and Language*, Volume 22, Issue 1, Pages 84-103, 2008.

[5] H. Satori, M. Harti and N. Chenfour. Introduction to Arabic Speech Recognition Using CMU SphinxSystem. *International Journal of Computer Science*, 2007.

[6] S. Seneff. TINA : a Natural Language System for Spoken Language Applications. *Computational Linguistics*, Volume 18, N°1, Pages 61–86, 1992.

[7] B. Van Schooten, S. Rosset, O. Galibert, A. Max, R. Op Den Akker and G. Illouz. Handling speech input in the Ritel QA dialogue system. *Proceedings of Interspeech'07*, Antwerp, Belgium, 2007.

[8] J. Villaneau. Une expérience de compréhension en contexte de dialogue avec le système LOGUS, approche logique de la compréhension de la langue orale. *TALN'07*, Toulouse, France, 2007.

[9] W. Ward. Extracting Information in Spontaneous Speech. In *Proceedings of International Conference of Speech and Language Processing, ICSLP*, Yokohama, 1994.

[10] A. Zouaghi, M. Zrigui and M. Ben Ahmed. Évaluation des performances d'un modèle de langage stochastique pour la compréhension de la parole arabe spontanée. *TALN'07*, Toulouse, France, 2007.

[11] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, J.T. Hazen and L. Hetherington. JUPITER: A telephone-based conversational interface for weather information. *IEEE Trans., on Speech and Audio Processing*, Volume 8, N°1, 2000.