Une ontologie pour le lexique arabe

DOUMI Noureddine
doumi@univ-saida.dz,
ndoumi@lycos.com

LEHIRECHE Ahmed
elhir@yahoo.com

Evolutionary Engineering and Distributed Information Systems Laboratory
(EDISL)
Département d'informatique, Université de Djilali Liabes, SBA, Algérie.

Résumé :
Dans le cadre du projet KalimNet lancé récemment dans le laboratoire EDISL – département d'informatique - université UDL de SBA, une proposition est envisagée de réaliser une base lexicale arabe sous forme d'ontologie. L'objectif est d'organiser le lexique arabe sous forme d'hiérarchie de concepts ce qui aboutira à un dictionnaire conceptuel exploitable par la machine (machine readable dictionary) de tels dictionnaires sont requis dans les systèmes de traitements automatique du langage naturel en tant que base de connaissances lexicales, et en général sont requis dans le projet Web sémantique afin de réaliser l'interopérabilité entre les agents logiciels des systèmes d'information sous web.
Des bases de données lexicales d'autres langues ont été déjà réalisées et commercialisées telle que WordNet réalisée à l'université de Princeton NJ-USA pour l'anglais, EuroWordNet pour les langues européennes.

Mots clés : Lexique, sémantique lexicale, relations sémantique, ontologie, base de connaissances linguistiques, systèmes de TALN.

1. Introduction

Afin de réaliser sa tache un système de compréhension automatique de langage naturel a besoin de deux composantes de connaissances : une base de connaissance lexicale et une base de connaissance du monde. La première sert comme une source de la sémantique lexicale des énoncés de la langue en question cette connaissance n'est pas suffisante tout seule pour l'accomplissement des taches de système, car il a besoin en plus des connaissances extralinguistiques, ces dernières sont structures dans une base de connaissances sous forme d'ontologie.
Différentes bases lexicales ont été réalisées pour des langues diverses, on cite à titre d'exemple WordNet de l'université de Princeton la plus connue et la plus réussite pour la langue anglaise, en d'autre part l'ontologie Cyc de D. Lenat et R. Guha représente la source de connaissances générales qui est généralement utilisée par les systèmes NLP [BAN2003].
Notre travail consiste à proposer une structure de base de connaissance pour le lexique arabe, en s'inspirant de WordNet notre travail consiste à exploiter la sémantique lexicale existante mais sous forme informelle -les dictionnaires traditionnels- et les convertir en une structure ontologique. Les sources adoptées par notre étude sont les dictionnaires arabes les plus conseils par les académiciens à savoir Lissane El Arab pour l'ancien arabe et El Mounjid pour l'arabe moderne.

2. La connaissance linguistique

Les deux principales composantes de la connaissance linguistique sont la connaissance lexicale et la connaissance syntaxique, le sens d'un énoncé est combiné par l'interaction des sens des unités lexicales, des mécanismes syntaxiques et du contexte de l'énoncé. La sémantique lexicale a pour objectif d'étudier les relations sémantiques entre les éléments du lexique d'une langue.

Relations sémantiques fondamentales

Les relations sémantiques qui vont être citées ici sont considérées comme étant les relations sémantiques fondamentales car elles forment la charpente de la structuration sémantique du lexique de toute langue. Chaque lexie se positionne dans le réseau lexical de la langue tout d'abord en fonction de ces relations.
Il s'agit des relations hyperonymie/hyponymie, meronymie/holonymie, antonymie, implication, voir aussi,...

3. La base lexicale WordNet

WordNet est une ontologie pour la langue anglaise basée sur des études psycholinguistiques et développée à l'université de Princeton par G. Miller. Cette base est considérée par quelques littératures d'ingénierie de connaissance comme une ontologie de haut niveau [TRM2001]. Elle a été conçue comme une ressource informatique qui couvre des catégories lexico-sémantiques appelées *synsets*. Les *synsets* sont des ensembles de synonymes qui regroupent des items lexicaux ayant des significations similaires comme par exemple les mots *``a board''* (un panneau) et *``a plank''* (une planche) groupés dans le *synset {board, plank}* [OLF2003]. Mais *``a board''* peut aussi désigner un groupe de personnes (un conseil d'administration par exemple) et pour désambiguïser ces significations homonymiques *``a board''* appartiendra aussi au *synset {board, committee}*. La définition des *synsets* varie du très spécifique au très général. Les *synsets* les plus spécifiques ne regroupent qu'un nombre restreint de significations lexicales alors que les *synsets* les plus généraux couvrent un nombre très large de significations.
L'organisation de WordNet à travers des significations lexicales au lieu d'utiliser des unités lexicales le rend différente des dictionnaires traditionnels et des thesaurus. L'autre différence que présente WordNet par rapport aux dictionnaires traditionnels se traduit par la séparation des données en quatre bases de données associées aux catégories de verbes, de noms, d'adjectifs et d'adverbes. Ce choix d'organisation est motivé par des recherches psycholinguistiques sur l'association de mots aux catégories syntaxiques par des sujets humains pendant des sondages psychologiques. Chaque base de données est organisée différemment des autres. Les noms sont organisés en hiérarchie, les verbes par des relations, les adjectifs et les adverbes par des hyper-espaces N-dimension.

4. L'ontologie

Le mot ontologie vient de la philosophie, d'un point de vue philosophique "Ontologie" est une branche philosophique qui s'occupe de la nature et l'organisation de la réalité. En informatique, les ontologies ont pour objectif de capturer la connaissance d'un domaine donné de façon générique et fournit une compréhension d'un commun accord de ce domaine qui peut être utilisée et partagée parmi les applications et les groupes [TRM2001]. Les ontologies fournissent le vocabulaire commun d'un domaine et définit le sens des termes –par les différents niveaux de formalité- et les relations entre eux.

On adopte la structure d'ontologie donnée par Maedche, selon ce dernier une ontologie consiste en un 5-tuplet composé des éléments essentiels d'une ontologie : concepts, relations, hiérarchie, une fonction qui relie les concepts non taxonomiquement et un ensemble d'axiomes [KJ2003].

$O := \{C, R, H, rel, A\}$

- Deux ensembles disjoints, C (concepts) et R (relations);
- Une hiérarchie de concepts H, c'est une relation dirigée, $H \subseteq C \times C$ appelée hiérarchie ou taxonomie de concepts, $H(c1,c2)$ veut dire que c1 est un sous concept de c2.
- Une fonction $rel : R \rightarrow C \times C$ reliant les concepts non taxonomiquement.
- Un ensemble d'axiomes ontologiques A exprimée en un langage logique approprié.

5. Les catégories syntaxiques arabes

Le lexique arabe est divisé en trois catégories syntaxiques principales qui sont les noms, verbes et prépositions, cette décomposition n'est pas adéquate à notre étude ce qui nous a obligé de chercher une autre décomposition qui permet de séparer les noms adjectivaux aux autres noms. Ainsi on a décomposé la catégorie nom en deux sous catégories (nom et adjectif).

6. Notre approche

L'objectif de notre travail est de réaliser une base lexicale arabe, sous forme d'ontologie linguistique capturant la connaissance linguistique arabe. On désigne ici par connaissance linguistique, les connaissance encodées en le lexique de la langue en question.

Pour atteindre notre objectif on s'est basé sur trois principes :

- On s'inspire de l'organisation de WordNet en tout ce qui concerne la structure ontologique et l'organisation de la connaissance linguistique;
- On utilise les dictionnaires traditionnels comme source riche de la sémantique lexicale arabe à savoir Lissane El Arab [LSN] et El Mounjid [MNJ1991];
- Le lexique étudié est choisi des titres des articles de sport des journaux arabes nationaux.

L'étude de WordNet (la catégorie nom) nous a permit d'observer que la profondeur de l'hiérarchie des synsets nominaux, ne dépasse pas la dizaine [WNN1993], et les synsets des niveaux les plus hauts sont des concepts d'abstraction de la cognition humaine donc peuvent être partagés par toute les langues et ne sont pas spécifique à la langue anglaise. Par conséquent on a traduit les 25 concepts top de la catégorie nom de WordNet en leurs équivalents arabes.

Exemple : {act, action, activity} → الحركة, {animal, fauna} → الحيوان, …, {group, collection} → التجمع …etc.

Les synsets résultats de la traduction servent comme des racines des hiérarchies de notre base.

Le reste de notre algorithme consiste en :

- Constituer les synsets (concepts);
- Trouver la position du synset (concept) dans l'hiérarchie;
- Détecter les relations lexicales et sémantiques autres que les relations taxonomiques.

Constitution des synsets

Le dictionnaire arabe est organisé autour des racines trilatérales et quadrilatérales, on applique des patrons sur ces racines pour dériver la majorité du lexique arabe, par conséquent pour trouver la signification d'un vocable arabe, on doit le réduire en sa racine (lemmatiser) et puis chercher son entée dans le dictionnaire. Le sens d'un mot est donné par soit : synonyme, antonyme ou une définition sous forme de genre spécifique suivi de différence spécifique.
Pour constituer le synset :

1. construire un synset d'un seul élément qui est le mot de corpus en question;
2. donner une clé numérique à ce nouveau synset comme celle de WordNet;
3. si le sens du mot est donné par un synonyme on ajoute les synonymes au synset;
4. on fait une fermeture transitive sur le nouveau synset pour trouver tous les synonymes possibles;

Exemple : تشكيلة l'entrée du mot indique que son sens est : مجموعة et le sens de ce dernier est : طائفة، جماعة، فريق en fin on constitue le synset {تشكيلة، مجموعة، طائفة، فريق}

La relation taxonomique

Le sens par fois est donné dans le dictionnaire par une définition sous forme de genre prochain plus les différences spécifiques.
Le genre prochain représente généralement l'hyperonyme immédiat du mot en question et ainsi peut être utilisé comme indicateur pour trouver la position du synset dans l'hiérarchie
Exemple: الفريق: الطائفة من الشيء المتفرق donc hyper(الفريق,الطائفة) c'est-à-dire le genre prochain de la définition الطائفة est un hyperonyme du mot en question الفريق.
Patron : le mot d'entête de la définition

Les relations non taxonomiques

Les différences spécifiques d'une définition expriment trois composantes de la connaissance lexicale: les parties, les attributs et les fonctions [WNN1993].
Les parties d'un concept nominal sont reliées à lui par la relation de meronymie/holonymie
Exemple : الزمرة: الفوج من الناس donc انسان qui est le singulier de ناس est une partie de زمرة donc on a la relation mero(زمرة,انسان)
Patron : tout les mots qui indique partie dans la définition tel que : من, أجزائه, له, يتكون من.
Les attributs sont des propriétés du concept nominal et sont exprimés sous forme d'adjectifs.
Exemple : الشنان: الماء البارد l'adjectif البارد est une propriété pour le concept الشنان donc on a la relation attr(البارد, الشنان).
Patron : tout adjectif dans la définition du mot en question est attribut.

Les fonctions représentent l'aspect fonctionnel du concept nominal, décrivent ce que les instances peuvent faire ou ce qui est fait par ou à ces instances. Les fonctions sont exprimées par les verbes cités dans une définition de dictionnaire.
Exemple : حبل يصعد به على النخلة : الراقول la fonction de l'entité désignée par الراقول c'est يصعد donc on a la relation fnct(الراقول,صعد).

Patron : tous les verbes d'action qui sont sous forme indéfinie (المبني للمجهول).

La base lexicale

L'opérationnalisation de la base est faite par une base de connaissance de format Prolog où la base des faits sont tous les tuples des relations lexicales et sémantiques et les règles sont les propriétés algébriques des relations.

5. Conclusion

La construction d'une ontologie de lexique est une tache qui demande la contribution des linguistes, des psycholinguistes et des ingénieurs de connaissances et en ce sens notre approche a essayé de bénéficier de l'effort des linguistes et psycholinguiste déployé dans la base lexicale la plus utilisée WordNet. Et vue que cette dernière en cours d'amélioration continue, on a profité d'ajouter quelques relations qui nous semblent porteuse de connaissances.

Bibliographie

[BAN2003]: Kornél Robert BANGHA, *La place des connaissances lexicales face aux connaissances du monde dans le processus d'interprétation des énoncés*, thèse de doctorat en linguistique et intelligence artificielle, Université de Montréal, Août 2003.

[BOU2002]: Caroline BOUSQUET-VERNHETTES, *Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique*, thèse de doctorat en informatique, Université de Toulouse III, Septembre 2002.

[JMP2000]: Jean-Marie Pierrel, *Ingénierie des langues*, HERMES Science publication, 2000

[MYR2004]: Myroslava O. Dzikovska, *A practical semantic representation for natural language parsing*, thèse de doctorat en informatique, Université de Rochester, New York, 2004.

[OLF2003]: Olfa Jenhani, *Ontologies pour le WEB: relations, construction d'ontologies et méthodes de raisonnement pour la génération de langue naturelle*, Rapport ARC GeNI, INRIA, Mai 2003.

[TRM2001]: Oscar Corcho, Mariano Fernández-López, Asunción Gómez Pérez, OntoWeb. D1.1. Technical Roadmap, Université polytechnique de Madrid, 2001.

[POL2002] : Alain POLGUÈRE, Notions de base en lexicologie, OLST, Université de Montréal, 2002

[WN1993]: George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller: Introduction to WordNet: An On-line Lexical Database, 1993

[WNN1993]: George A. Miller: Nouns in WordNet: A Lexical Inheritance System, 1993

[WNA1993]: Christiane Fellbaum, Derek Gross, and Katherine Miller: Adjectives in WordNet, 1993

[KJ2003]: Karin Koogan Breitman, Julio Cesar Sampaio do Prado Leite: Lexicon Based Ontology Construction: 2003

[MNJ1991]: El Mounjid en langue et célébrités, 31eme édition, Dar El-Machreq SARL publishers, Bierut - Liban, 1991

[LSN]: Iben El Moundhir, Lissane El Arab, version électronique.