Enrichissement sémantique du corpus iSPEDAL

Abd El Salam AL HAJJAR
Mohammad HAJJAR
Zeinab ABDEL NABI
Georges LEBBOS
Laboratoire GRIT
Institut Universitaire de Technologie
Université Libanaise, Liban

Résumé : Dans cet article nous présentons la méthodologie utilisée pour doter iSPEDAL d'une dimension sémantique. iSPEDAL est une version améliorée de DESELA, peut être présenté sous la forme d'une base de données relationnelle facilement exploitable à l'aide des langages de requêtes appropriés. Dans notre cas trois voies sont explorées pour ajouter de certaines relations de sens entre les mots dans iSPEDAL. La première est basée sur l'exploitation des caractéristiques naturelles d'un dictionnaire classique, c'est qu'un dictionnaire propose, en générale, pour un mot donné ses synonymes, ses antonymes, etc. La deuxième voie est basée sur la traduction vers et à partir de l'anglais pour projeter les relations sémantiques entre les mots anglais aux mots arabes. La dernière consiste à exploiter le « Word-Net arabe » pour enrichir notre dictionnaire.

Mot-clé : Langage arabe, dictionnaire arabe, dimension sémantique, translation, WordNet arabe.

1. Introduction

La proposition des dictionnaires classiques arabes a constitué l'essentiel des travaux linguistiques effectués sur la langue arabe. La plupart de ces dictionnaires sont maintenant disponibles sur le web sous forme des fichiers électroniques plats. Donc, on observe actuellement, de plus en plus, une transition vers des dictionnaires électroniques [Al Hajjar et al., 2010]. Par contre, la plupart de ces dictionnaires collecte leurs données à partir des plusieurs dictionnaires classiques et offre un service de navigation et de recherche limité. Ces limites d'interrogation sont dues, principalement, à une faiblesse de structuration des entrées dictionnairiques utilisées [Habash, 2005] [Habash, 2004] [Habash and

Rambow, 2006]. Dans ce cadre, nous avons proposé un vrai dictionnaire électronique structuré et évolutif de la langue arabe iSPEDAL [Hajjar et al., 2010] qui est une version améliorée de DESELA [Al Hajjar et al., 2009a]. En effet, iSPEDAL peut être présenté sous la forme d'une base de données relationnelle facilement exploitable à l'aide des langages de requêtes appropriés. Ce nouveau dictionnaire fournit les liens d'un mot donné avec sa racine, ses affixes et son modèle éventuel. De plus, nous avons construit un système automatique qui permet d'alimenter et d'enrichir iSPEDAL à partir de plusieurs dictionnaires classiques ou à partir d'un corpus textuel arabe quelconque. Par contre, iSPEDAL souffre d'une handicape par rapport à un dictionnaire classique et qu'il n'offre pas des relations sémantique entre ses mots [Al Hajjar, 2010].
Dans cet article nous présentons la méthodologie utilisée pour doter iSPEDAL d'une dimension sémantique qui se restreint à l'ajout de certaines relations de sens entre les mots. Dans notre cas trois voies sont explorées. La première est basée sur l'exploitation des caractéristiques naturelles d'un dictionnaire classique, c'est qu'un dictionnaire propose, en générale, pour un mot donné ses synonymes, ses antonymes, etc. La deuxième voie est basée sur la traduction vers et à partir de l'anglais pour projeter les relations sémantiques entre les mots anglais aux mots arabes. La dernière consiste à exploiter le « Word-Net arabe » pour enrichir notre dictionnaire.

2. Le dictionnaire électronique arabe iSPEDAL

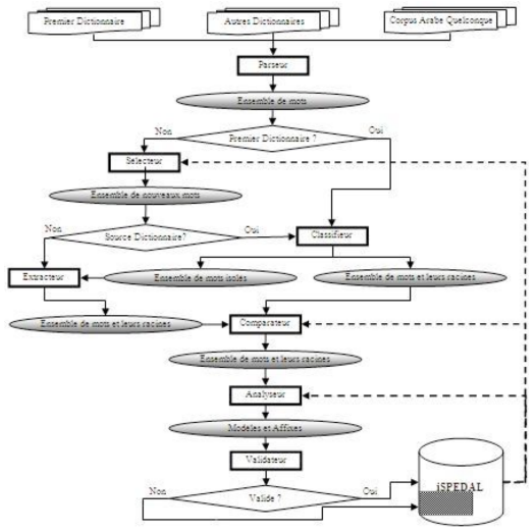

Figure 1 : Schéma général d'iSPEDAL

La figure 1 présente le Schéma général selon lequel iSPEDAL fonctionne. Elle montre l'architecture générale du système d'alimentation et d'enrichissement automatique à partir d'un ou de plusieurs dictionnaires classiques ainsi qu'à partir des corpus textuels arabes quelconques. iSPEDAL est constitué de plusieurs composantes qui sont : le Parseur, le Sélecteur, le Classifieur, l'Extracteur, le Comparateur, l'Analyseur et le Validateur.

La première composante de ce système est le parseur qui permet de transformer le document en un ensemble des mots, selon les séparateurs qui sont généralement des espaces.

L'entrée de ce système peut être un dictionnaire arabe classique sous format plat (fichier texte, PDF,..) ou n'importe quel autre corpus arabe textuel (page web, fichier texte, …) [Al Hajjar et al., 2010]. Si le document en entrée est le premier dictionnaire, l'ensemble des mots est passé au classifieur. Ce premier dictionnaire est utilisé pour initialiser iSPEDAL [Hajjar et al., 2010]. Dans les autres cas, c'est le sélecteur qui reçoit l'ensemble des mots.

Le rôle du sélecteur est d'éviter les doublons en s'assurant que les mots à ajouter à iSPEDAL n'y sont pas déjà. La sortie de cette composante est un ensemble des nouveaux mots qui est soumit au classifieur, si cet ensemble est en provenance d'un dictionnaire, ou à l'extracteur dans le cas contraire. Le classifieur est la composante qui permet de scinder l'ensemble des mots reçus en entrée, en deux sous ensemble : d'un coté les racines et leurs mots dérivés qui sont envoyés vers le comparateur, d'un autre les mots isolés qui sont envoyés vers l'extracteur.

Cette séparation est basée sur le format du dictionnaire d'entrée, où les racines sont encadrées par des séparateurs spéciaux et les mots, qui sont situés après cette racine et avant la racine suivante, dérivent de la première. L'extracteur utilise la méthode d'extraction détaillée dans pour trouver la racine d'un mot arabe [Al Hajjar et al., 2009b].

Les ensembles des mots associés à leurs racines, en provenance du classifier et de l'extracteur, sont soumis au comparateur qui permet d'éviter les doublons, à tous les niveaux, dans iSPEDAL [Hajjar et al., 2010].

L'ensemble des nouveaux mots et des racines associées sont utilisés par l'analyseur pour produire les affixes et les modèles. La sortie de cette composante est un ensemble des mots, des racines, des modèles et des affixes.

Ces ensembles sont soumis au validateur pour approuver ces résultats ainsi que les liens entre eux. Le validateur utilise les éléments essentiels de la morphologie de la langue arabe pour l'approbation de ces éléments. Seuls les éléments valides sont ajoutés à iSPEDAL, le reste est mis dans une zone tampon en attente d'une validation ultérieure [Al Hajjar, 2010].

3. La dimension sémantique dans iSPEDAL

Dans cet article nous limitons la dimension sémantique à l'ajout de certaines relations de sens entre les mots. Ces relations permettent de situer le descripteur dans son environnement conceptuel [Walde and Zinsmeister, 2006]. Ces relations sont :

Synonymie : est un rapport de similarité sémantique entre des mots ou des expressions d'une langue. La similarité sémantique indique qu'ils ont des significations très semblables. Des termes liés par synonymie sont des synonymes [Pagin, 2000] [Walde and Zinsmeister, 2006] [Berzlánovich et al., 2008].

Antonymie : Deux mots sont en relation d'antonymie si on peut montrer une symétrie de leurs traits sémantiques par rapport à un axe. Ils sont de sens contraire [Mohammad et al., 2008] [Walde and Zinsmeister,2006] [Berzlánovich et al.,2008].

Hyponymie : est la relation sémantique d'un lexème à un autre selon laquelle l'extension du premier est incluse dans l'extension du second. Le premier terme est dit hyponyme de l'autre. C'est le contraire de l'hyperonymie [Berzlánovich et al., 2008] [Walde and Zinsmeister,2006].

Hyperonymie : est la relation sémantique hiérarchique d'un lexème à un autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. Le premier terme est dit hyperonyme de l'autre, ou super ordonné par rapport à l'autre [Berzlánovich et al., 2008] [Walde and Zinsmeister,2006].

Méronymie : est une relation partitive hiérarchisée, une relation de partie à tout. Des termes liés par méronymie sont des méronymes. Un méronyme A d'un mot B est un mot dont le signifié désigne une sous-partie du signifié de B. La relation inverse est l'holonymie. Par exemple, اليد est un méronyme de الجسم, de même que سقف est un méronyme de البيت [Berzlánovich et al., 2008] [Walde and Zinsmeister,2006].

Holonymie : est une relation partitive hiérarchisée, c'est la relation inverse de la relation méronymie. Des termes liés par holonymie sont des holonymes. Un holonyme A d'un mot B est un mot dont le signifié désigne un ensemble comprenant le signifié de B. Par exemple, الجسم est un holonyme de اليد, البيت est un holonyme de سقف [Berzlánovich et al., 2008] [Walde and Zinsmeister, 2006].

4. Méthodologie de mise en place des liens sémantiques dans iSPEDAL

Pour ajouter les relations de sens entre les mots à iSPEDAL, nous avons utilisé trois méthodes.

La première est basée sur l'exploitation des caractéristiques naturelles d'un dictionnaire classique (Fig. 2). En effet, un tel dictionnaire propose, en générale, pour un mot donné la définition, l'orthographe, les sens, les synonymes, les antonymes, les modes d'utilisation, etc. Notre procédure consiste à trouver les mots clés qui signifient qu'après ces mots on peut trouver des autres mots ou des phrases qui ont des relations sémantiques avec le mot ou la racine en question. Par exemple, dans le dictionnaire « Lesan Al Arabe » [Ibn Manzour, 2008], on peut trouver sous la racine «الكل» le mot clé «أي» qui permet de trouver le synonyme de cette racine («طعام»). Donc, il faut recenser les expressions qui expriment des relations sémantiques avec l'élément d'entrée du dictionnaire. D'autre part, on peut trouver sous chaque racine, des expressions qui expliquent les mots dérivés d'elle. De plus, on peut trouver des expressions qui donnent quelques caractéristiques de ces mots (pluriel, antonyme,…).

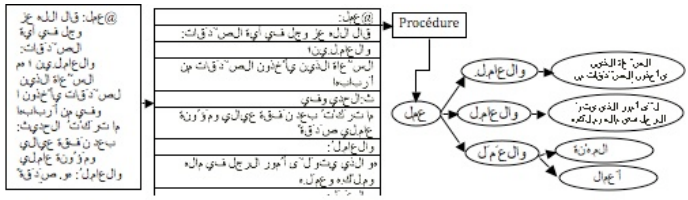

Figure 2 : Extraction des relations entre les mots à partir du dictionnaire « Lesan Al Arabe ».

La deuxième voie est basée sur la traduction vers et à partir de l'anglais pour projeter les relations sémantiques entre les mots anglais aux mots arabes (Fig. 3). Il y a beaucoup des traducteurs arabes [Sakher, 2010] [Diab, 2004], anglais ou français. Si deux mots arabes ont un même sens, probablement ils ont la même traduction en anglais, ou bien en français. Pour cela, nous prenons 2 mots arabes, on les traduit, la traduction de chaque mot peut donner un ensemble de mots, donc nous avons 2 ensembles des mots A et B, s'il y a une intersection entre ces 2 ensembles alors il y a une relation sémantique entre les 2 mots d'entrée avec un coefficient de ressemblance CR qui est définit par $A \cap B / A \cup B$. Exemple :

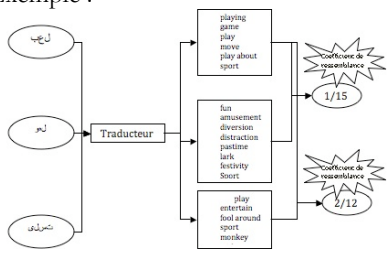

Figure 3 : Extraction des relations entre les mots en utilisant la traduction.

La dernière consiste à exploiter le « Word-Net arabe » pour enrichir notre dictionnaire. Word-Net est une grande base de données lexicale qui classe les mots arabes en noms, verbes, adjectifs, adverbes, etc. [Miller et al., 1993] , [Rennie, 2000] ,[Pedersen et al. 2004](Fig. 4). Ces éléments sont regroupés en ensembles de synonymes cognitifs (synsets) dont chacun exprime un concept distinct. Ces synsets sont reliés entre eux par des relations conceptuelles-sémantique et lexicales. La relation principale entre les mots dans Word-Net est la synonymie, par exemple : (إغلاق et إقفال) [Abouenour, 2008], [Abouenour, 2010], [Bouzoubaa, 2010]. Notre procédure consiste à extraire ces ensembles et à projeter les relations au niveau des mots de chaque ensemble.

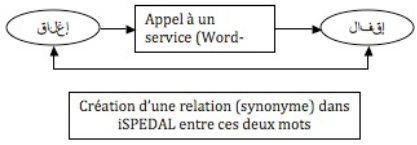

Figure 4 : Extraction des relations entre les mots à partir du Word-Net Arabe.

Bibliographie

[Abouenour et al.,, 2008] L. ABOUENOUR, K. BOUZOUBAA, P. ROSSO. Improving Q/A Using Arabic Wordnet. In: Proc. The 2008 International Arab Conference on Information Technology (ACIT'2008),Tunisia, December. 2008.

[Abouenour et al.,, 2010] L. ABOUENOUR, K. BOUZOUBAA, P. ROSSO. Using the YAGO ontology as a resource for the enrichment of Named Entities in Arabic WordNet. Workshop LR & HLT for semitic languages, LREC'10. Malta. May 2010.

[Al Hajjar et al., 2009a] A. AL HAJJAR, M. HAJJAR, K. ZREIK "Un nouveau dictionnaire électronique structuré et évolutif pour la langue arabe", CiDE.12, 12e Colloque International sur le Document Electronique, Montréal, Canada, 21 - 23 Octobre, 2009.

[Al Hajjar et al., 2009b] A. AL HAJJAR, M. HAJJAR, K. ZREIK "Classification of Arabic Information Extraction methods" MEDAR 2009 2nd International Conference on Arabic Language Resources and Tools, Le Caire, Egypte, 21-23 Avril, 2009.

[Al Hajjar et al., 2010] A. AL HAJJAR, M. HAJJAR, K. ZREIK "Structure, historique et évolution des dictionnaires arabes : le cas d'iSPEDAL", CiDE.13, 13e Colloque International sur le Document Electronique, Paris, France, 16-17 Décembre, 2010.

[Al Hajjar, 2010] A. AL HAJJAR "Extraction et gestion de connaissance à partir du Web multilingue : Spécificité de la langue arabe", Université Paris 8, France, Co-directeurs Pr. K. Zreik et Pr. M. Hajjar, thèse de doctorat soutenue le 17 décembre 2010.

[Berzlánovich et al., 2008] I. BERZLÁNOVICH, Lexical cohesion and the organization of discourse, University of Groningen, Supervisors: G. Redeker, M. Egg., 2008.
[Bouzoubaa, 2010] K. Bouzoubaa, Arabic Wordnet Use and Enrichment, Mohammadia School of Engineers, Rabat, Morocco. 2010.
[Diab, 2004] M. DIAB, Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet, Linguistics Department Margaret Jacks Hall Stanford University Stanford, CA 94305, USA.2004.
[GT, 2011] Google Traduction (GT), web site : http://translate.google.fr/.2011.
[Habash and Rambow, 2006] N. HABASH and O. RAMBOW, A Morphological Analyzer and Generator for the Arabic Dialects, Center for Computational Learning Systems Columbia University New York, NY 10115, USA. july 2006.
[Habash, 2004] N. HABASH, Large Scale Lexeme Based Arabic Morphological Generation, University of Maryland Institute for Advanced Computer Studies University of Maryland College Park College Park, Maryland, 20742 USA. 2004.
[Habash, 2005] N. HABASH, Arabic Natural Language Processing: Words, Columbia University Center for Computational Learning Systems, Summer School on Human Language Technology Johns Hopkins University, Baltimore July 6th, 2005.
[Hajjar et al., 2010] M. HAJJAR, A. AL HAJJAR, K. ZREIK, P. GALLINARI "A improved structured and progressive electronic dictionary for the Arab language: iSPEDAL", IEEE ICIW 2010, The Fifth International Conference on Internet and Web Applications and Services, Barcelone, Espagne, 9 - 15 Mai, 2010.
[Ibn Manzour, 2008] IBN MANZOUR, 2008. Lisan Al-Arab, www.muhaddith.org.
[Miller et al., 1993] G. MILLER, R. BECKWITH, C. FELLBAUM, D. GROSS, K. MILLER, Introduction to WordNet: An On-line Lexical Database, 1993
[Mohammad et al., 2008] S. MOHAMMAD, B. DORR, G. HIRST, Towards Antonymy-Aware Natural Language Applications, University of Toronto, 2008.
[Pagin, 2000] P. PAGIN, A Quinean definition of synonymy, 17 May 2000.
[Pedersen et al., 2004] T. PEDERSEN, S. PATWARDHAN, J. MICHELIZZI. WordNet: : Similarity - Measuring the Relatedness of Concepts. In: AAAI (2004), p. 1024-1025. 2004.
[Rennie, 2000] RENNIE, J. WordNet::QueryData: a Perl module for accessing the WordNet database. http://search.cpan.org/dist/WordNet-QueryData. 2000.
[Sakher, 2010] Sakher Company, Al Ajeeb, http://lexicons.ajeeb.com 1998-2010.
[Walde and Zinsmeister, 2006] S. WALDE, H. ZINSMEISTER, Introduction to Corpus Resources, Annotation and Access: Semantic Annotation, Foundational Course 18th European Summer School in Logic, Language and Information Málaga, Spain July 31 - August 4, 2006.