# A Hybrid Approach for Arabic Semantic Relation Extraction

**Wiem Lahbib[1], Ibrahim Bounhas[2], Bilel Elayeb[3,4], Fabrice Evrard[5], Yahya Slimani[6]**

[1] LISI Laboratory of computer science for industrial systems,
Carthage University, Tunisia
e-mail: wiemlahbib88@hotmail.fr
[2] LISI Laboratory of computer science for industrial systems,
Higher Institute of Documentation, La Manouba University, Tunisia
e-mail:Bounhas.Ibrahim@gmail.com
[3] RIADI Laboratory, The National School of Computer Science, Manouba University, Tunisia.
[4] Emirates College of Technology, Millennium Tower, Sheikh Hamdan Street, P.O. Box: 41009.
Abu Dhabi, United Arab Emirates.
e-mail: Bilel.Elayeb@riadi.rnu.tn
[5] Informatics Research Institute of Toulouse, 02 Rue de Charles Camichel, Toulouse, France.
e-mail: Fabrice.Evrard@enseeiht.fr
[6] LISI Laboratory of computer science for industrial systems,
Higher Institute of Multimedia Arts of Manouba,
Manouba Universtity, Tunisia
e-mail:Yahya.Slimani@gmail.com

**Abstract**

Information retrieval applications are essential tools to manage the huge amount of information in the Web. Ontologies have great importance in these applications. The idea here is that several data belonging to a domain of interest are represented and related semantically in the ontology, which can help to navigate, manage and reuse these data. Despite of the growing need of ontology, only few works were interested in Arabic language. Indeed, arabic texts are highly ambiguous, especially when diacritics are absent. Besides, existent works does not cover all the types of semantic relations, which are useful to structure Arabic ontologies. A lot of work has been done on cooccurrence- based techniques, which lead to over-generation.

In this paper, we propose a new approach for Arabic semantic relation extraction. We use vocalized texts to reduce ambiguities and propose a new distributional approach for similarity calculus, which is compared to cooccurrence.

We discuss our contribution through experimental results and propose some perspectives for future research.

**Keywords**: Arabic ontology, Morphological and Syntactic Analysis, Semantic relation extraction

## Introduction

Information retrieval systems (IRS) are developed to automate and facilitate the information access. Ontologies support these systems by allowing semantic and navigation-based search. Thus, it will be possible to rapidly access to relevant information with concepts used explicitly to describe and represent a knowledge domain. Building an otology requires the extraction of concepts and relationships from a specified domain. Nowadays, all semantic relations existing between concepts can be represented thanks to ontologies. However, the extraction process based on Arabic textual resources is not yet well investigated; compared to other languages. Arabic knowledge extraction is a complex task, because the written code of Arabic is an ambiguous one. The existence of diacritics can certainly minimize the level of complexity and facilitate the analysis tasks. That's why, we chose to trait vocalized Arabic corpora hopping to overcome the ambiguity.

In this paper, we present a hybrid approach which extracts noun phrases at the first stage, and then, transforms them into semantic relations.

This document is organized as follows. The second section deals with a literature review. The third one is about the proposed method to extract syntactic and semantic relations form Arabic texts. In the fourth one, we explain the choice of the test corpora. Our experiments and results will be detailed and interpreted in the fifth section. Finally, and before concluding, we will make the differences, strengths and weaknesses of our approach compared to related works.

## Related works

Semantic knowledge integrated in IRS has many origins and its extraction from texts can be based on different approaches. We introduce an overview on principal existing

approaches for semantic relation extraction from Arabic textual corpora.

**Clustering-based approaches:** The main idea of clustering is to propose classification methods of data. For example, the auto-expansion method adopted by (Pinto, 2007) can be considered as one of clustering approaches. We should mention that calculating co-occurrence frequencies depends on the definition of context as basis of co-occurrence. The problem arises when the terms are set in the same part of the text but they are not necessarily semantically close. So it brings us to the over-generation. A possible solution is to use other types of contexts. For instance, using distributional analysis based on contextual dependencies showed successful results (Bounhas et al., 2011b).

**Derivation-based approaches:** The approach of Belkredim and El Sebai (2009) is based on derivation and inflexion rules. The author proposed to link verbs with their derivatives. The ontology so, will be constituted by derived nouns belonging to six categories (verbal noun, active participial, etc.) and derived verbs. This approach causes over-generation of noise.

**Mapping-based approaches:** To build the Arabic WordNet (AWN), (Elkateb, 2006) presented a mapping based-approach. Once the most important concepts are represented to create the core of the ontology, arabic terms must be linked to their English corresponding entries. This mapping allows to enrich ontology's concepts. It requires the use of ontologies in other languages (i.e. English). For example, in AWN project the mapping was project (Aliane, 2010) used GOLD ontology. Jarrar (2011) proposed to map knowledge extracted from machine readable dictionaries and linked to the concepts of SUMO and DOLCE with the English WordNet. This type of approach is limited, because we should guarantee the translation process and accuracy.

**Pattern-based approaches**: This type of approach was adopted by Bouzoubaa (2011) to enrich AWN. First, we need to identify all hyperonymy patterns which can be found in the web. Then, we must find hyponym/hyperonym couples for each pattern. Often, this type of approach gives promising results, but the choice of patterns must be done carefully, in order to obtain reusable and high-coverage patterns. By studying these approaches, we may conclude by the following remarks:

- Related work in Arabic ontology construction considered many types of relations. However, some types of relations have not yet been studied. We mainly talk about non-taxonomic relations, such as possession, causality, etc. which may be transformed from syntactic relations (cf. figure 1).
- Existent work in arabic ontology learning focused on cooccurrence which proved to cause over-generation (Bounhas et al., 2011b).
- Pattern-based approaches impose rigid constraints and do not take into account the fuzziness and ambiguity of semantic relations. That is, the correspondence between syntactic and semantic relations is ambiguous and fuzzy.

That is, our goals in this paper are: i) to take into account new types of relations (cf. figure 1); ii) to propose syntactic dependencies as mean to model context while computing similarity; and, iii) to model fuzziness in relation extraction by adopting an enrichment of signatures.

## The Proposed approach

In this paper we suggest a new approach for semantic relation extraction from Arabic texts. It's a hybrid approach which mixes statistical calculus and linguistic knowledge, thus taking advantage of the benefits and the limits of these techniques. The idea is to exploit syntactic dependencies to infer semantic relations. Compared to other state-of-the art approaches (e.g. Vetulani, 2004) which focuses on verbs, we exploit nous phrases as are the most meaningful entities (Malaisé et al., 2003). Our approach uses linguistic tools to analyze texts and establish correspondence between syntactic dependencies and semantic relations. It also uses statistical measures to compute the similarity between terms and interpret syntactic relations. Semantic relation extraction must consider three analysis levels:

**Morphological analysis**: this step allows analyzing the words of a given corpus and identifying their attributes. We mainly talk about POS (grammatical category), definiteness, gender, number and stem. These attributes are used in the syntactic analysis step, where we are interested in nouns, adjectives and prepositions.

**Syntactic analysis**: it is the second preparatory step in this work. During this step, we aim at extracting all noun phrases (NP) present in an Arabic text. Based on the studies presented by bounhas and slimani (2009) and Bounhas et al. (2011a), we distinguish five types of noun phrases: adjective phrases (مركب نعتي), prepositional phrases ( مركب حرفي), annexation phrases (مركب إضافي), conjunctive phrases (مركب عطفي) and complex noun phrases. In each NP, the first component represents the head and the second is the expansion. This task is relatively less complicated with vocalized texts. We have developed a parser which matches between two axes: the morphological information and the rules of Arabic grammar. It is an approach that is not limited to the identification of a noun phrase as is the case of existing parsers of Arab corpus (Boulaknadel, 2008), but our analyzer allows to recognize the syntactic tree of each phrase and identify its constituent's roles. As a result, we generate a syntactic network linking heads and expansions with different types of syntactic relations.

**Semantic analysis**: The main idea behind our approach can be represented by a graph as a summary of dif-

ferent relationships which may be inferred from a prepositional phrase (cf. figure 1)[1].

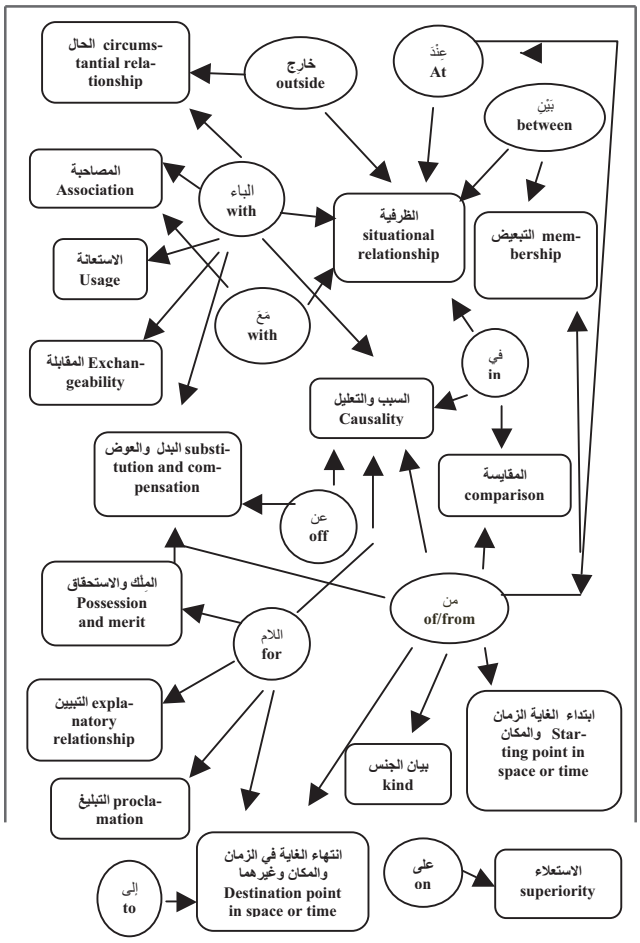


***Figure 1.*** *Correspondence between prepositional and semantic relations*[1]

We represent the most common prepositions in our corpus and we show that each one reflect one or more semantic relations. It should be noted that rules of Arabic grammar are behind this relationship between semantics and syntactic relations.

As we can see some cases have ambiguities in the extraction of relations. This brings us to propose an answer to this kind of ambiguity, called enrichment of signature based on distributional analysis. If we consider (t1, t2) a pair of terms, the signature is defined by the set C= {c1, c2, ..., cn} such as $c_i$, is a syntactic relation between t1 and t2. Our disambiguation approach tries to enrich the signature of an ambiguous couple (t1, t2) by adding the signature of the nearest couple (t1', t2'), where t1' (respectively t2') is the closest term to t1 (respectively t2). To compute the similarity, we use contingency table based measures (Pazienza et al., 2005). This table is a square matrix composed of four values (O11, O12, O21, and O22). We propose to asses two different definitions of these values

## The proposed architecture

Our system presented by figure 2, has been designed to extract noun phrases, store them in a database in order to extract semantic relations. It consists of five modules.

**Morphological analyzer:** We chose to integrate the AraMorph tool (Buckwalter, 2002). It is an analyzer developed in Java and works on the transliteration of the Arabic word and it has a level of ambiguity of 02.60. It gives as a result the token, its morphosyntactic category and its English translation. The choice of this tool is justified by its ease of integration and the fact that it is designed for classical Arabic texts considered in our experiments.

**Syntactic analyzer:** This module applies grammar rules corresponding to each type of noun phrase. It analyses the file provided by AraMorph and checks each attribute useful in parsing. Expressions which apply to these rules are stored in the database.

**Similarity calculus module**: The calculus of similarity helps in identifying the meaning behind a set of terms considered close by these calculations. (Elkhlifi, 2011) used this technique to extract opinion-oriented words and to classify the polarity of general opinion. Regarding our approach, we use the similarity calculus to allow cases enrichment technique. To calculate the distances between the terms we use several measures among which we mention T-score (TS), LLR score (i.e. Likelihood Ratio»), DF («Dice Factor») and MI («Mutual Information») (Pazienza et al., 2005). The scores are calculated as follows:

$$\mathrm{LLR}(u,v) = -2\log\left(\frac{L(O_{11},C_1,r)*L(O_{12},C_2,r)}{L(O_{11},C_1,r_1)*L(O_{12},C_2,r_2)}\right)$$

where

$$L(k,n,r) = k^r * (1-r)^{n-k}$$

$$r = R_1/N, r_1 = O_{11}/C_1, r_2 = O_{12}/C_2$$

$$MI = \log_2(O_{11}/E_{11})$$

$$TS = \frac{(O_{11}-E_{11})}{\sqrt{O_{11}}}$$

$$DF = 2*\frac{O_{11}}{R_1+C_1}$$

**Enrichment module:** This module uses different distances generated by the module for similarity calculus to initiate the process of case enrichment. It firstly retrieves couples and enriches their cases. We also assigned weights to relationships on noun phrases as their order of appear-

[1] This dependency-graph has been built based on a study of arabic grammar books, mainly the book intitled Al-Osol fi Al-Nahw of Abou Bakr Al-Saraj , which died in 316 H.
(الأصول في النحو لأبي بكر ابن السراج).

ance in the enrichment process. Thus, the signature added in the first iteration has a greater weight then the one added in the second iteration, and so on.

**Validation module:** This module compares the semantic relations extracted using a given type of distance to a reference list provided by an expert, thus computing a success rate.

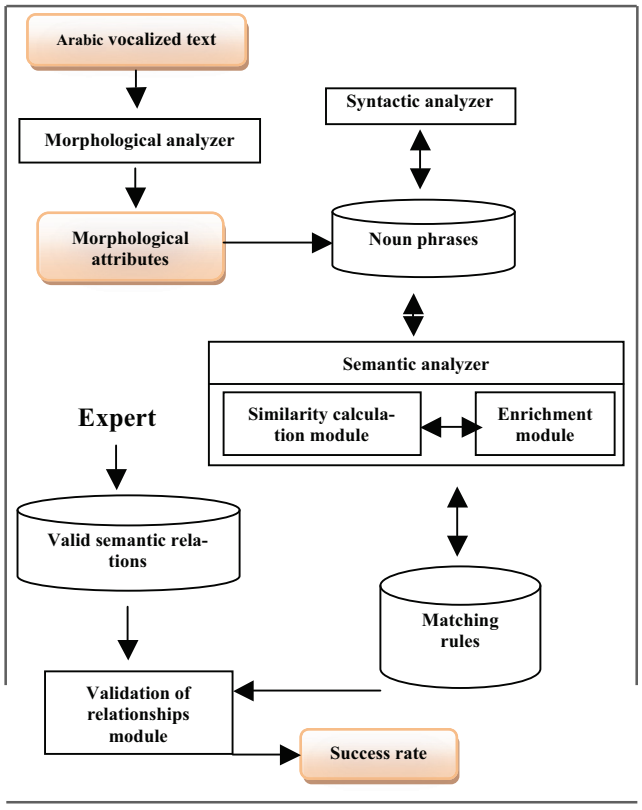


***Figure 2.*** *The system's architecture*

## Test Corpora

We choose a vocalized corpus to avoid different types of ambiguities resulting from the lack of diacritics. The corpus of hadith is the best candidate for the extraction of semantic relations for many reasons. First, the multitude of application areas shows the importance of this corpus (Bounhas et al., 2011ab; Hasanain, 2009; Zaraket and Makhlouta, 2012). In addition, the structure of hadith documents is rich of socio-semantic knowledge and in terms of noun phrases. Finally, several versions of the corpus are vocalized. It is also possible to apply our approach to other types of corpus since we use linguistic and statistical tools which are applicable to any type of Arabic. We performed experiments in different domains. Three areas are considered: drinks, purification and fasting. Table 1 presents the characteristics of each of the three areas of our selected test corpus.

| Area | Number of word | Number of distinct noun phrases |
|---|---|---|
| Drinks | 2540 | 480 |
| Purification | 2800 | 370 |
| Fasting | 2510 | 433 |

***Table 1.*** *Characteristics of the hadith by area*

## Experiments and results

In this section, we present the results obtained by our approach. In the following sections, we present results for different table of contingency table setup and compare our results to the state-of-the-art co-occurrence-based approach.

### Results with the first definition of the contingency table

Table 2 summarizes the results obtained in the three considered areas. The lines show similarity scores and the columns represent the different thresholds of these scores. Each cell contains the percentage of correctly extracted semantic relations.

On this evaluated sample, we note that the success rates of semantic relation extraction presented by these three domains are weak and do not exceed 65%. This problem can be explained by a partial failure in the operation of the signature enrichment. We noted that this process does not affect the entire heads-expansions couples constituting the phrases. We are faced with situations where a term in a given NP has no expansions in other phrases. Thus, we perform the same experiments with the second definition of the contingency table (i.e. computing both common expansions and common heads).

| Drinks | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.1 |
| LLR | 65 | 65 | 65 | 65 | 50 | 50.10 | 50.1 | 44.6 |
| MI | 65 | 65 | 65 | 47 | 47 | 46.8 | 57.44 | 57.4 |
| TS | 65 | 65 | 65 | 65 | 47 | 46.8 | 57.88 | 57.8 |
| DF | 6 5 | 65 | 65 | 65 | 65 | 65 | 65 | 46.8 |
| Purification | | | | | | | | |
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.2 | 0.1 |
| LLR | 49.54 | 49.54 | 49.54 | 49.54 | 49.54 | 49.54 | 49.54 | 49.54 |
| MI | 52.25 | 52.25 | 52.25 | 52.25 | 52.25 | 52.25 | 52.25 | 40 |
| TS | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| DF | 53.15 | 53.15 | 53.15 | 53.15 | 53.15 | 53.15 | 53.15 | 53.15 |
| Fasting | | | | | | | | |
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.2 | 0.1 |
| LLR | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 |
| MI | 60.68 | 60.68 | 60.68 | 60.68 | 54.9 | 54.9 | 54.9 | 54.9 |
| TS | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60 |
| DF | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 | 60.68 |

***Table 2.*** *Success rates of relations extraction with the first definition*

### Results with the second definition of the contingency table

This new definition has undoubtedly improved outcomes. As table 3 shows, the maximum percentage exceeded 70% and reached 75% with a small decrease of 0.02% in DF score.

| Drinks | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.2 | 0.1 |
| LLR | 75 | 75 | 75 | 75 | 50 | 50 | 50 | 50 |
| MI | 75 | 75 | 75 | 75 | 75 | 75 | 75 | 70 |
| TS | 75 | 75 | 75 | 75 | 75 | 75 | 75 | 72 |
| DF | 77 | 77 | 77 | 77 | 77 | 77 | 77 | 77 |
| Purification | | | | | | | | |
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.2 | 0.1 |
| LLR | 66 | 66 | 66 | 66 | 66 | 66 | 66 | 60 |
| MI | 66 | 66 | 66 | 66 | 66 | 66 | 66 | 59 |
| TS | 66 | 66 | 66 | 66 | 60 | 60 | 60 | 60 |
| DF | 69 | 69 | 69 | 69 | 69 | 69 | 69 | 69 |
| Fasting | | | | | | | | |
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.1 |
| LLR | 70 | 70 | 70 | 70 | 70 | 70 | 70 | 70 |
| MI | 96.9 | 96.9 | 96.9 | 96.9 | 96.9 | 96.9 | 96.9 | 96.9 |
| TS | 59 | 59 | 59 | 59 | 59 | 59 | 59 | 59 |
| DF | 70 | 70 | 70 | 70 | 70 | 70 | 70 | 70 |

***Table 3.*** *Success rate of relationships extraction: second description*

However, by analyzing in detail our results, we remarked that many signatures have not been enriched what explains failures in semantic relation extraction. To illustrate this fact, we present in table 4, the results obtained for couples whose signatures where enriched to assess the impact of signature enrichment. The results of this experiment show high success rates compared with results from the first two tables.

Indeed, the maximum rate increased for the three areas to 97% and 100% in the field of purification, by the LLR score.

| Drinks | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.2 | 0.1 |
| LLR | 94.7 | 94.7 | 94.7 | 94.7 | 94.7 | 94.7 | 94.7 | 70 |
| MI | 95.9 | 95.9 | 95.9 | 95.9 | 95.9 | 95.9 | 95.9 | 73 |
| TS | 85.5 | 95.9 | 95.9 | 95.9 | 60 | 60 | 60 | 60 |
| DF | 95.9 | 95.9 | 95.9 | 95.9 | 95.9 | 95.9 | 85 | 85 |
| Purification | | | | | | | | |
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.2 | 0.1 |
| LLR | 100 | 100 | 100 | 100 | 66 | 66 | 60.7 | 60 |
| MI | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 76 |
| TS | 90 | 90 | 90 | 90 | 90 | 90 | 90 | 81 |
| DF | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 55 |
| Fasting | | | | | | | | |
| | 1 | 0.9 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.1 |
| LLR | 97 | 97 | 97 | 97 | 97 | 97 | 97 | 97 |
| MI | 95 | 95 | 95 | 95 | 95 | 95 | 81 | 81 |
| TS | 95 | 95 | 95 | 95 | 95 | 95 | 95 | 95 |
| DF | 97 | 97 | 97 | 97 | 97 | 97 | 97 | 97 |

***Table 4.*** *Success rate of relationships extraction: second description improved*

As an example, we take this prepositional phrase (Drinking in a receptacle). By applying our approach, we have succeeded to extract the right semantic relation which is (situational relationship).

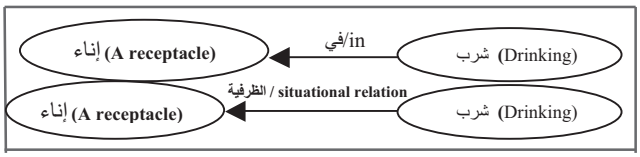


***Figure 3.*** *Syntactic relationship transformed on semantic relationship*

**Comparison with the co-occurrence-based approach:** we compared our results to the co-occurrence-based approach classically used in the field of arabic document indexing and knowledge extraction with those presented by our approach. Table 5 shows that we experiment different values of co-occurrence threshold (i.e. the minimum number of times where the two terms co-occur).

| | 10 | 20 | 40 |
|---|---|---|---|
| Drinks | 55.00% | 60.00% | 60.00% |
| Purification | 67.42% | 59.20% | 40.00% |
| Fasting | 57.00% | 57.00% | 55.00% |

***Table 5.*** *Success rate of relationships extraction: Co-occurrence*

From this table, we note that the co-occurrence method gives results close to 50%, almost the half of the relationships are not distinguished. In the first hand, experiments have shown the contribution of our approach relatively to the co-occurrence. We note that our approach, assuming only the situations where the cases enrichment is carried out, is more effective than the technique of co-occurrence. We managed to correctly extract all relations in the field of purification and we recorded 60% as the most decreased rate, on the other side, co-occurrence reaches 67% as the best result. In the second hand, we remark that these two approaches extract the same relationship in some cases. However, in many other cases, we remarked that the two approaches are complementary.

### Synthesis

By analyzing these results, we may remark that:

- Syntactic parsing has a great role in our approach, since it influences the enrichment process whose impact was clearly shown.
- Our experiments showed that all the similarity scores (LLR, MI, TS and DF) reached close success rates.
- The results reveal a complementarily of syntactic dependencies with co-occurrence linking.

• Our approach reaches the best rates with the drinks area. Indeed, this domain contains more nous phrases, thus the signature of a couple of terms have greater chance to be enriched.

## Conclusion

In this paper, we presented and evaluated a new approach for semantic relation extraction from Arabic corpora. We used a vocalized corpus to reduce ambiguities. We used the hadith corpora because the existence of structured vocalized versions. Our experiments showed the contribution of signature enrichment based on syntactic dependency study compared to co-occurrence linking, but also revealed the complementarily of these two approaches. This may be considered in future work covering bigger data collections. We also plan to enlarge the scope of our work by considering other types of compound nouns like annexation or adjectival phrases. We will also try to compare to the state-of-the art approaches like contextual exploration (Alrahabi et al., 2006). Finally, no work is interested, as we know, in this type of relationship that we cannot compare our work to other work.

## References


Aliane, H., Alimazighi, Z., Mazari, A. 2010. Al -Khalil : The Arabic Linguistic Ontology Project In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 17-23 May 2010.

Alrahabi, M., Ibrahim Amr, H. , Desclés J-P. 2006. Semantic Annotation of Reported Information in Arabic, FLAIRS 2006, Melbourne, Florida, May 11-13, 263-268

Belkredim F., El Sebai A. 2009. An Ontology Based Formalism for the Arabic Language Using Verbs and their Derivatives, Communications of the IBIMA, Vol: 11(5), Issue: 5.

Boulaknadel S., Daille B., and Aboutajdine D. 2008. A multi-word term extraction program for arabic language. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), pages 1485-1488, Marakech, Morocco, May 17-23 2008.

Bounhas, I. and Slimani, Y. 2009. A hybrid approach for arabic multi-word term extraction. In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE), pages 429-436, Dalian, China, August 21-23 2009.

Bounhas I., Elayeb B., Evrard F., and Slimani Y. 2011a. Organizing contextual knowledge for arabic text disambiguation and terminology extraction. Knowledge Organization, 38(6), pp. 473-490.

Bounhas I., Elayeb B., Evrard F. and Slimani Y. 2011b. Arabonto: Experimenting a new distributional approach for building arabic ontological resources. International Journal of Metadata, Semantics and Ontologies (IJMSO), 6(2), pp. 81 - 95.

Bouzoubaa, K. 2011. Extending AWN with nouns and verbs and realizing a web prototype, In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Tunis, July 26-28 2011.

Buckwalter, T. 2002. Buckwalter Arabic Morphological Analyzer version 1.0. Linguistic Data Consortium, University of Pennsylvania.

URL : http ://www.buckwalterframing.com

Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C. 2006. Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

Elkhlifi, A., Bouchlaghem, R. and Faiz, R. 2011. Opinion Extraction and Classification Based on Semantic Similarities: Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press.

Hasanain, T.Z. 2009. Automatic Question Answering System for Arabic Language Textual Data, Mémoire de mastère, Faculty of Computing and Information Technology,Saudi Arabia.

Jarrar, M. 2011. Building a Formal Arabic Ontology Methodology and Progress. In proceedings of the Experts Meeting On Arabic Ontologies And Semantic Networks. Alecso, Arab League. Tunis, July 26-28 2011.

Malaisé, V., Zweigenbaum, P. and Bachimont, B. 2003. Vers une combinaison de méthodologies pour la structuration de termes en corpus : Premier pas vers des ontologies dédiées à l'indexation de documents audiovisuels. In Widad Mustafa El Hadi ed., Actes du 4e Congrès ISKO France 3-4 july 2003 Grenoble France, Paris, France : L'Harmattan, pp. 179-189.

Pazienza, M., Pennacchiotti, M. and Zanzotto, F. 2005. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches, In Spiros Sirmakessis eds., Knowledge Mining Series: Studies in Fuzziness and Soft Computing, Berlin, Heidelberg: Springer, pp. 255–279.

Pinto, D. and Rosso, P. and Benajiba, Y. and Ahachad, A. 2007. Word Sense Induction in the Arabic Language: A Self-Term Expansion Based Approach. In Proceedings 7th Conference on Language Engineering, The Egyptian Society Of Language Engineering, ESOLE.

Vetulani Z. 2004. Towards a Linguistically Motivated Ontology of Motion: Situation Based Synsets of Motion Verbs. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA. AAAI Press.

Zaraket F. and Makhlouta J. 2012. Arabic Cross-Document NLP for the Hadith and Biography Literature: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida. May 23-25, 2012. AAAI Press