Towards an Arabic Ontology
Defining Morpho-lexical Patterns for Semantic Relation Extraction

Mohamed Mahdi Boudabous[1], Fatiha Sadat[2] and Lamia Hadrich Belhuith[1]
[1]*ANLP Research Group, Laboratoire MIRACL, University of Sfax, Sfax, Tunisia*
[2]*Université du Québec à Montréal (UQAM), Montréal, Canada*

Keywords: Arabic Ontology, Morpho-lexical Patterns, Wikipedia, Semantic Relations, Pattern Definition.

Abstract: In this paper, we propose a method for defining morpho-lexical patterns used to detect semantic relation between Arabic nouns. This method is based on study corpus built from online encyclopedia. This corpus consists of a set of articles selected on the basis of a database containing pairs of terms linked by semantic relations. Defined patterns are then implemented using NooJ platform. The pattern evaluation result is very encouraging. We obtained 79% as F-Measure rate.

1 INTRODUCTION

Over the last decade, with the important development of the Web, information sources have become multiform and very rich. These sources are represented in different forms, understandable by users but not by computers. So the need of techniques and tools for automatic pre-processing information, which allows the computer to understand the information and transfer the content to the humans, is becoming important now days.

Ontologies are amongst the most powerful knowledge representation tools for modeling and managing various applications ranging over Natural Language Processing (NLP), information retrieval, semantic Web, etc (Baccar, 2012). In this context, research on ontology building has become increasingly widespread in computer science community.

Recently, ontologies have emerged as a major research topic in the fields of automatic language processing, information retrieval and semantic web (Baccar, 2012).

Several methodologies for building ontology have emerged over the past scores of years namely:

- From scratch methodology;
- Re-engineering methodology;
- Cooperative construction methodology;
- Integration methodology;
- Learning methodology.

Learning methodology differs according to data sources used for learning (Ben Mustapha, 2006) such as texts, dictionaries, knowledge bases, relational schemas and semi-structured schemas.

Through the analysis of the state of the art of learning methods, we can identify a methodological framework that consists of four steps:

- Corpus construction;
- Linguistic analysis of the corpus;
- Semantic standardization;
- Development of the operational ontology.

In this paper, we focus on the linguistic analysis of the corpus which allows the passage from linguistic level to conceptual one. In fact, we propose a pattern based method for semantic relation detection between the ontology concepts. The originality of this method consists in the definition of morpho-lexical patterns from Arabic wikipedia. Defined patterns allow then to extract semantic relations between nouns.

The present paper is outlined as follows: the first section is an introduction as we have seen above. The second section is devoted to present the basic concepts of domain ontology building. The third section presents an overview of approaches for extracting semantic relations. The fourth section exposes our method for defining morpho-lexical patterns from online encyclopedia and gives details on the corresponding stages. In the fifth section, we present the implementation of defined patterns, obtained results and discussion. The last section presents the conclusion and the prospects of our work.

Mahdi Boudabous M., Sadat F. and Hadrich Belguith L..
Towards an Arabic Ontology - Defining Morpho-lexical Patterns for Semantic Relation Extraction.
DOI: 10.5220/0004157603420348
In *Proceedings of the 14th International Conference on Enterprise Information Systems* (SCOE-2012), pages 342-348
ISBN: 978-989-8565-11-2

2 BASIC CONCEPTS

The concept of an ontology is inherited from a philosophical tradition that focuses on the science of Being. Over the last two decades, several definitions and types of ontologies have emerged since the introduction of this concept in computer science domain (Arara, 2002). In this section, we review the various definitions, the components of ontology, types of ontologies and methodologies for building ontologies.

In literature, many definitions are proposed for the concept ontology. The most referenced and synthetic one is that of Gurber (Gurber, 1993), which predicts that an ontology is "an explicit specification of conceptualization". This definition is expanded in 1997 by Borst (Borst, 1997), then by Studer and al. in 1998 (Studer, 1998) that define ontology as "an explicit and formal specification of a shared conceptualization".

According to Gurber (Gurber, 1993), ontologies are composed of five components namely concepts, relationships, properties, axioms and instances. The concepts represent an object in the universe. The relations reflect the relevant links existing between the concepts in the field of study. The properties are attributes that characterize the concepts and relationships. The axioms are used to constraint the value of concepts and relations. The instances are used for representing elements in a domain.

After building the first ontologies, the researchers define dimensions to classify ontologies. In fact, different dimensions of classification have emerged. The most known is the dimension proposed by Gomez-Pérez in 2004 (Gómez-Pérez, 2004) which classifies ontologies according to their objects of conceptualization. The types of ontologies best known along this dimension are:

- Generic ontologies (top-level ontologies): this type contain general concepts common to all domains or multiple domains (time, space, object, event);
- Domain ontologies: this type contains a set of vocabularies and concepts that describe an application domain;
- Tasks ontologies: this type is used to conceptualize specific tasks in the system;
- Application ontologies: this type is used to define ontologies that depend on both the domain and the task.

In the ontology engineering field, many methodologies are proposed for building ontologies. In fact, the first methodology for building ontology is called “from-scratch”, aimed to design a process of building ontologies in the absence of knowledge (Ben Mustapha, 2006). Given the limitations presented by this methodology, the researchers propose ontology re-engineering methodology that tends to bind ontology under implementation to another already built. After that, researchers propose new methodology that follows a collaborative approach including the intervention of people located in different places. This methodology called “cooperative construction methodology”. Moreover, with the deployment and the diversification of electronic resources, the researchers opt for the learning methodology which is based on heterogeneous data sources.

Building ontologies from texts is a sub-domain of engineering ontologies. Actually, several methods are involved for learning ontologies from texts. These methods differ according to the techniques used for extracting concepts and relations. Based on this criterion, these methods are grouped into three families of approach namely: statistical approaches, linguistic approaches and hybrid approaches (Ben Mustapha, 2006).

In this paper, we are interested in learning methods from texts based on linguistic techniques. Most of the existing methods are based on linguistic techniques. Added to that, these methods treat indo-European languages (George, 1993), (Vossen, 1998) and use lexico-syntactic patterns to detect semantic relations (Laignel, 2011). Moreover, to our knowledge there is no research works that are done on Arabic lexicon ontology for nouns. However, there are some Arabic ontologies among them we can cite Arabic WordNet (Black, 2006), Amine Arabic WordNet (Abouenour, 2008) and Arabic Ontology (Jarrar, 2011).

However, given the lack of tools for parsing Arabic language and the complexity of treatments specific to that language, we propose to define morpho-lexical patterns to identify semantic relations used to build a lexical ontology for Arabic.

3 OVERVIEW OF SEMANTIC RELATIONS EXTRACTION METHODS

In this section, we present an overview of the different techniques for extracting semantic relations.

Automatic identification of semantic relations in text is a difficult problem, although it is important for many applications (Green, 2002). Thus, different methods for extracting semantic relations from texts

have been proposed. These methods can be grouped into three main categories namely: statistical techniques, linguistic techniques and hybrid techniques.

Statistical techniques are based on the principle that terms which co-occur together are strongly linked by semantic relations. These techniques use statistical methods to calculate the distribution of words in the corpus (Agirre, 2000) or probabilistic methods to calculate the probability of occurrences of a set of terms (Velardi, 2001), (Neshatian, 2004). Once the concepts are detected, the relations that connect them can be identified by calculating the similarity between their syntactic contexts (Hindle, 1990), (Grefenstette, 1994), using either Bayesian networks (Weissenbacher, 2007) or text mining techniques (Grcar, 2007) or learning algorithms (Giuliano, 2006).

However, the major drawback of these methods is that they don't always identify the correct semantics of the relationship, so they require human intervention (Kergosien, 2009).

Linguistic techniques are based on the structure of sentences and texts. These techniques require automatic processing tools for texts analyses such as segmentation tools and grammatical tagging tools. Moreover, they are based on lexico-syntactic patterns to recognize linguistic markers of semantic relations (Aussenac-Gilles, 2000), or on contextual rules based on index and triggers.

The application of these techniques is carried out at the sentence, while other studies analyzed all level of the text. Thus, several methods have been implemented for example CHAMELEON (Aussenac-Gilles, 2000) and SEEK (Jouis, 1994). This type of approach is also used by the Edelweiss (Khelif, 2006).

Compared to numerical techniques, these techniques have a major advantage view that is the ability to identify semantics of the relationship.

Likewise, hybrid techniques combine statistical methods and linguistic methods. They generally use the syntactic distribution of terms to extract relations.

4 THE PROPOSED METHOD FOR SEMANTIC RELATION EXTRACTION

We propose in this section a method for extracting semantic relations from texts in order to build a lexical ontology from online encyclopedia (Fig. 1). Our method consists on defining a set of morpho-lexical patterns specific to each relationship that will be used to extract the relationships between concepts.

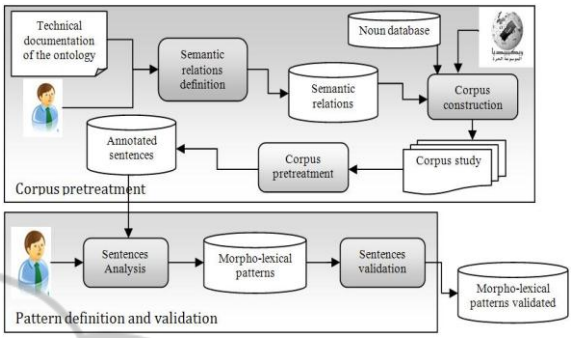

Figure 1: The two phases of the proposed method.

Figure 1 shows the two phases of the proposed method for defining semantic relationships between concepts. The first phase consists on the pre-treatment of the corpus and the second one consists on the definition and validation of patterns. Defined patterns ensure the passage from the textual level to the conceptual one and form the informal to the formal.

In the following sections, we detail the various steps of this method.

4.1 Corpus Pre-treatment

In this section, we detail the different steps of the pre-treatment corpus phase. The first step tends to define semantic relationships specific to the domain of the ontology. The second step consists on building the corpus and it is crucial and delicate. The third step is the corpus pre-treatment (segmentation, sentences extraction and morphological analysis).

4.1.1 Semantic Relations Definition

The definition of semantic relationships between concepts of an ontology is performed by an expert based on the technical documentation of the ontology. This documentation allows us to describe the characteristics of the ontology, such as the main objectives, the type of the ontology and application areas.

In this work, we aim to build a lexical ontology for Arabic language. This ontology must contain Arabic lexicons and semantic relations between them and it will be useful for several applications of NLP such as Information Retrieval, Question Answering (Q/A), etc.

From this description, the domain expert defines semantic relationships. In our work, the domain

expert defines eleven semantic relationships. Table 1 presents some examples of the defined semantic relations.

Table 1: Some examples of the defined semantic relations.

Semantic relation	Example	
Hyponymy (IS_A)	Animal	حيوان
	« is hyponym of »	
	Dog, horse	كلب، حصان
Hyperonymy	Cat	قط
	« is hyperonym of »	
	Animal	حيوان
Holonymy	Car	سيارة
	« is holonym of »	
	Door	باب،
Meronymy	Motor	محرك
	« is meronim of »	
	Car	سيارة
Member	France	فرنسا
	« is member of »	
	Union European	الاتحاد الاروبي
Synonymy	Person	شخص
	« is synonym of »	
	Individual	فرد
…		…
Antonymy	Father	أب
	« is antonym of »	
	Mother	أم

4.1.2 Corpus Construction

The proposed method for detecting semantic relationships is based on a corpus study. This corpus is representative as it contains indicative sentences. An indicative sentence is a sentence in the article that indicates a semantic relationship between two terms.

In fact, in the first step we build a database composed of pairs of concepts from Arabic WordNet (AWN). This database consists of 8000 nouns connected by 13,000 relationships extracted from AWN and enriched by the expert. In the second step we use the online encyclopedia Arabic Wikipedia to download articles corresponding for the list of database nouns. The choice of Arabic Wikipedia is justified by the fact that it is currently the largest source of knowledge on the web (i.e. it contains more than 160 000 articles in Arabic).

The acquisition of the articles constituting our study corpus is done in an automatic way. Our corpus consists on 2050 articles (Downloaded in March 2012).

4.1.3 Corpus Pre-treatment

The following automatic processing tasks are applied to the corpus in order to be able to define patterns:

- Segmentation: allows to segment articles in sentences. Segmentation is based on the punctuation markers, coordination conjunctions and some alternative keywords (Belguith, 2005);
- Extracting indicative sentences: this step tends to extract indicative sentences from articles;
- Sentences annotation: it consists of the sentences morphological analysis in order to determine for each unit (i.e. a word or a compound word) the part of speech, gender, number, tense, etc.

At the end of this phase all indicative sentences are morphologically annotated. These sentences will serve as an input for the patterns definition and validation phase.

4.2 Patterns Definition and Validation

In this section, we detail the steps of the morpho-lexical patterns definition phase. We define a morpho-lexical pattern as a linguistic structure or schema that consists of a set of words and / or morphological categories in a specific order. The first step groups sentences that have specific semantic relationship. The second step has a manual scan of all sentences belonging to the same relationship in order to extract patterns related to each relationship. The third step allows validating defined patterns. These patterns are defined to automate the process of detecting semantic relations from texts.

4.2.1 Grouping of Sentences

This step categorizes sentences into groups according to the semantic relations they indicate. This task is done with reference to the database of nouns defined in the first phase. Then, we group sentences that have a common morphological structure.

The main interest of this step is to facilitate the task of defining patterns specific to each relationship.

4.2.2 Sentences Analysis

In this state, the domain expert is responsible for the study of morphological structures of sentences in order to define patterns. This study tends to extract a

morpho-lexical pattern from each sentence or group of sentences. Defined patterns consist in morphological categories of words and lexicons. These lexicons are used to specify words that indicate the relationship. Table 2 shows some examples of sentences and the morpho-lexical pattern matching.

Table 2: Examples of indicative sentences and morpho-lexical pattern matching.

Phrases	Morpho-lexical pattern
يتكون الهواء من غاز النيتروجين The Air is composed of nitrogen gas	(اسم معرف) + أداة (حرف جر) + X + مضارع فعل + (اسم معرف)Yاسم نكرة + Present verb + X (Noun) + Tool (preposition) + Noun + Y (Noun)
يتكون الماء من ذرتي هيدروجين The water is composed of two atoms of hydrogen	

4.2.3 Pattern Validation

After defining the patterns, it is often useful to validate the list of patterns by a human expert. This step tends on conducting a comparative study between all patterns of different relationships to ensure that each pattern is specific to a relationship.

Based on this comparative study, we note the existence of some ambiguous patterns, that is to say belonging to more than a relationship. To resolve this problem we add lexical indicators to differentiate ambiguity patterns (table 3).

Table 3: Examples of an ambiguous pattern.

Relations	Phrases	Morpho-lexical patterns	Solution
Antonymy	الحزن هو عكس السعادة Grief is the opposite of happiness	X (اسم معرف) + ضمير فصل (هو، هي) + اسم نكرة + Y (اسم معرف) X (Noun) + Pronoun (He, she) + Noun + Y (Noun)	اسم نكرة: عكس، مقابل، نقيض. Noun: opposite, reverse, versus.
synonymy	النقد هو شكل المال Currency is the form of money	X (اسم معرف) + ضمير فصل (هون، هي) + اسم نكرة + Y (اسم معرف) X (Noun) + Pronoun (He, she) + Noun + Y (Noun)	اسم نكرة: شكل، نموذج، مرادف. Noun: synonymous, form, model.

Table 3 shows two indicative sentences that have the same morphological structure but belong to two different relations. In this case we need a mark to distinguish the two patterns. For this example the solution proposed by the domain expert is to define the list of prepositions possible specific to each relationship.

5 PATTERNS IMPLEMENTATION

Once patterns are validated by the domain expert, the next step consists in implementing all defined patterns. For this, we use NooJ platform. This platform allows us to represent specific patterns to each relationship as a grammar (Figure 2). In fact, we characterize morpho-lexical elements present before, after and between related terms. Also, we characterize the form of the elements related by markers. These elements are shown in parentheses in the pattern. For example, in Figure 2, we use the marker "var" to characterize the first term and "var1" to characterize the second term.

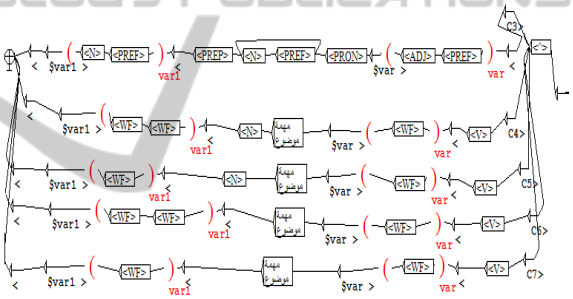

Figure 2: Example of a NooJ grammar.

Obtained grammars are combined into a single grammar called “main grammar” in order to be able to apply all the grammar on all sentences of the corpus to extract all relations of the ontology.

In order to evaluate the effectiveness of defined patterns on the recognition of semantic relations, we apply the principal grammar on a test corpus. This corpus consists of 300 articles extracted from the online encyclopedia Arabic Wikipedia. These articles contain 370 relationships.

The first step consists in segmenting articles into sentences. Then, in the second step, sentences are annotated morphologically in order to apply defined patterns and detect semantic relations in an automatic way.

By comparing pairs of words and semantic relationships that link them automatically extracts with those extracted by the domain expert, we can calculate measures of recall, precision and F-measure for each relationship. In fact, we determine,

for each pair of words detected, the relevance of the terms and the semantic relationship between them. Two values are possible:

- The value "valid": if the relationship is deemed valid and the two words linked by this relationship;
- The value "false": when the relationship is wrong (in this case, the terms are often invalid).

Obtained results are shown in Table 4.

Table 4: Evaluation results of morpho-lexical patterns.

Relations	Recall	Precision	F-measure
Average	78%	85%	79%

Following the evaluation of built grammars, we obtain values of recall, precision and F-measure, respectively: 78%, 85% and 79%. Although the results are encouraging, there are relationships that are not detected by the defined patterns. By analyzing non-detecting pairs of words, we find that this problem is mainly due to two reasons:

- The lack of patterns that cannot recognize the relationship: In fact, in some cases the patterns, concerning some phrases, are not defined;
- The non-detection of compound nouns: using NooJ morphological analyzer, we notice that it is unable to detect compound nouns, which prevents from applying some patterns to the sentences.

Moreover, in some cases we notice that the patterns recognize erroneous relationships. This can be explained by two reasons:

- The existence of some ambiguous patterns;
- Morphological analysis using NooJ platform produce several morphological categories: in some cases NooJ return more than one morphological category of a word, which allow to us apply several patterns on one sentence and detect more than one semantic relation between two terms.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a method to define morpho-lexical patterns that is useful for extracting semantic relations between arabic nouns. This method is based on a corpus of 2050 Arabic Wikipedia articles. We have detailed the various steps of the proposed method. As a result, the defined patterns are implemented using NooJ platform in order to automatically identify the pairs of words and the semantic relations between these specific terms. Nevertheless, our method uses a minimum of knowledge, based primarily on morpho-lexical knowledge, obtained results are very encouraging (i.e. 78% recall, 85% precision and 79% F-measure) which proves the importance of morpho-lexical patterns in the detection of semantic relations for Arabic language.

As future perspectives, we plan to resolve ambiguous patterns. In addition, we intend to propose a method for detecting compound nouns. Finally, we consider applying these patterns to build a lexical ontology for Arabic.

REFERENCES

Baccar, F., Gargouri, B., Ben Hamadou, A., 2012. Towards Generation of Domain Ontology from LMF Standardized Dictionaries. *In Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE)*.

Ben Mustapha, N., Aufaure, M. A., Baazaoui Zghal, H., 2006. Vers une approche de construction de composants ontologiques pour le Web sémantique – synthèse et discussion. *In Troisième atelier Fouille de Données Complexes dans un processus d'extraction des connaissances.*

Arara, A., Ben Slimane, D., 2002. Towards ontologies building: a terminology based approach, *In VLDB Journal - Special Issue on Semantic Web*.

Gurber, T. R., 1999. Toward Principles for the Design of ontologiesused for Knowledge Sharing. *In International Journal of Human-Computer Studies. Vol. 43 , pp. 907-928.*

Borst, W., 1997. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. *Ph.D. Dissertation.*

Studer, R., Benjamins, V. R., Fensel, D., 1998. Knowledge Engineering: Principles and Methods. *In Data Knowl. Eng., vol. 25, pp. 161–197.*

Gómez-Pérez, A., Fernández-López, M., Corcho, O., 2004. Ontological engineering: with examples from the areas of knowledge Management, e-Commerce and the semantic Web.

George, A. M., Beckwith, R., Fellbaum, C., Gross, D., Miller K., 1993. Introduction to wordnet: An On-line Lexical Database. *In International Journal of Lexicography*.

Vossen, P., 1998. Eurowordnet: A Multilingual Database with Lexical Semantic Networks. *Livre*.

Laignel, M., Kamel, M., Aussenac-Gilles, N., 2011. Enrichir la notion de patron par la prise en compte de la structure textuelle - Application à la construction d'ontologie. *In TALN 2011 (Traitement automatique des langues naturelles).*

Black, W., Elkateb, A., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Fellbaum, C., 2006. Introducing the Arabic wordnet Project.

Abouenour, L., Bouzoubaa, K., Rosso, P., 2008. Construction de l'ontologie Amine Arabic wordnet dans le cadre des systèmes Q/R. *In Proceedings de la Journée Scienfique sur les Technologies de l'Information et de la Communication, JOSTIC 08*.

Jarrar, M., 2011. Building a Formal Arabic Ontology. *In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*.

Green, R., Bean, C. A., Myaeng, S. H., 2002. The Semantics of Relationships An Interdisciplinary Perspective, *Springer*.

Agirre, E., Ansa, O., Arregi, X., Artola, X., Diaz, A., Lersundi, M., Martinez, D., Sarasola, K., Urizar, R., 2000. Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar.

Velardi, P., Missikoff, M., Basili, R., 2001 Identification of relevant terms to support the construction of Domain Ontologies. *In: ACL-EACL Workshop on Human Language Technologies, Toulouse, France.*

Neshatian, K., Hejazi M. R., 2004. Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies. *In: 2nd Workshop on Information Technology and its Disciplines. Pp. 43–48.*

Hindle, D., 1990. Noun classification from predicate argument structures. *In: proceeding of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90). Berkeley USA.*

Grefenstette, G., 1994. Explorations in Automatic Thesaurus Discovery. *MA: Kluwer Academic Plubisher. Boston.*

Weissenbacher, D., Nazarenko A., 2007. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. *In TALN conférence*.

Grcar, M., Klein, E., Novak, B., 2007. Using Term-Matching Algorithms for the Annotation of Geo-services. *In Postproceedings of the ECML-PKDD 2007 Workshops.*

Giuliano, C., Lavelli, A., Romano, L., 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. *In Proceeding EACL*.

Kergosien, E., Kamel, M., Sallaberry, C., Bessagnet, M. N., Aussenac, N., Gaio, M., 2009. Construction automatique d'ontologie et enrichissement à partir de ressources externes. *In 3es Journées Francophones sur les Ontologies.*

Aussenac-Gilles, N., Seguela, P., 2000. Les relations sémantiques : du linguistique au formel. *In Numéro spécial linguistique de corpus. Toulouse.*

Jouis, C., 1994. SEEK, un logicielle d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. *In Actes des 6ème Journées Acquisition, Validation. INRIA, pp. 159—172.*

Khelif, K., 2006. Web sémantique et mémoire d'expériences pour l'analyse du transcriptome. *These de doctorat.*

Belguith Hadrich, L., Baccour, L., Ghassan, M., 2006. Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. *In 12ème conférence sur le Traitement Automatique des Langues Naturelles.*