Automatic construction of ontology from Arabic texts

Ahmed Cherif Mazari[1] , Hassina Aliane[2], and Zaia Alimazighi[3]

[1]Electrical Engineering and Computer science Department, University of Médéa
mazari.ac@gmail.com
[2]CERIST, Research Center on Scientific and technical Information, Algiers.
haliane@mail.cerist.dz
[3]Computer Science Department, USTHB, Algiers.
alimazighi@wissal.dz

Abstract. The work which will be presented in this paper is related to the building of an ontology of domain for the Arabic linguistics. We propose an approach of automatic construction that is using statistical techniques to extract elements of ontology from Arabic texts. Among these techniques we use two; the first is the "repeated segment" to identify the relevant terms that denote the concepts associated with the domain and the second is the "co-occurrence" to link these new concepts extracted to the ontology by hierarchical or non-hierarchical relations. The processing is done on a corpus of Arabic texts formed and prepared in advance.

Keywords: Ontology, Information Extraction (IE), Arabic Natural Language Processing (Arabic-NLP), Statistical methods for text processing.

1 Introduction

Existing methods of ontologies construction differ mainly according to the information that they treat (concepts, relations, properties ...) and techniques for extracting these elements from texts. These techniques are carried out either by methods that require linguistic corpus annotated or by statistical methods that do not need the annotation text. In our approach, we are oriented toward the use statistical methods, since these methods do not require these types of annotated corpora and NLP[1] analyzers (such as the lexical analyzer and parser). These methods are based on two criteria: the relevance of a term from a domain that is defined by the number of occurrences of the word in the corpus and the co-occurrence of two terms at a frequency more high.

2 Overview of the Approach

In our approach, we started the initialization of the ontology manually, by the general (generic) concepts retrieved from the ontology of GOLD (General Ontology for Linguistic Description) [Far03], it is a general ontology for descriptive linguistics and is applicable to most human languages. It was created on the base of the general

[1] NLP: Natural Processing Language.

ontology of SUMO[2] (the Standard Upper Merged Ontology). Then, we adopted the process of extraction from the domain text which can be summarized in three main steps; the first is the formation of the domain corpus, this step is fundamental since the quality of the corpus will depend on the quality of processing and the corpus must fully cover the domain treated. The second step is the extraction of candidate terms (*these terms may be among the elements that make up the ontology: a concept, a relation or an individual*). Finally, we make the junction of these new elements to the ontology.

2.1 Constitution and preparation of the corpus

In a project of construction ontologies from texts, the corpus, its status and its collection are of paramount importance both as a source of knowledge to build the model and also a source of reference throughout the process development [BoA03]. So the questions addressed in the constitution of the corpus include: the type of corpus (a corpus "specialized" is a corpus containing texts on a topic related to a domain of knowledge as our case Arabic linguistics), and the suitability for the project referred (the quality of the results of a corpus largely is depending on the quality of the corpus, this means, that the domain texts are well defined and delimited, they are fairly representative). However, size is often limited by the availability of texts and issues of copyright). Representativeness (variety of texts, authors, sources, etc) and using full-texts or samples. [Mar03]

Preparation of corpus. After the formation of crude corpus, it must be prepared for processing. This phase is performed by a set of preprocessing steps to remove some ambiguity, reduce the number of transactions and adapt the corpus following the final objective “extraction of candidate terms”.

Normalization. In the corpus, we will encounter elements that do not carry information and increase the processing time. This is mostly special characters, numbers, non-Arabic words, abbreviations and single letters. These should be deleted:

- Special characters: include any special sequence of characters delimited by letters or spaces.
- Numbers: We regroup all the character sequences located between two spaces containing numbers in a single occurrence. This method also has the advantage to combine the dates, the actual numbers and percentages.
- Words in Latin characters: The non-Arabic words, mainly in Latin characters are simply detected by their graphic.
- Abbreviations and isolated letters: The list of words to a single letter in the Arabic texts reveals the presence of a significant number of these words. These letters are often used in abbreviations. It may designate a variable, for example ب الفئة«, category B », numbering ; أ الفقرة « section A », ت for تاريخ «date», م for ميلادي,

[2] http://suo.ieee.org developed in the project *IEEE SUO* Working Group.

ص for صفحة «page». We can find also letters that form a grammatical category for example (ا ، و ، ي) :حروف العلة. [AbD08]

- Character '–':The typographers make frequent use of the character **'-',** allowing the extension of the line in the middle of words, for better readability, to limit the white space on a line justified, even for purely aesthetic reasons. This character is not part of the Arabic alphabet. It is therefore necessary to eliminate it.
- To remove the vowel signs, which are written in the form of diacritics placed above or below letters.
- Because of graphs variations that may exist when writing the same word and that they can be sources of ambiguity. We will make some substitutions as follows: Substituting letters إ, آ and أ by ا. Substituting of end letters ي , ة by ى,ه . [Dou05]

Deletion of Stop-Words. These are grammatical or lexical words; they are so often grouped together in a "stop-list." It is generally accepted that these words very common (about half of the occurrences of a text) are not indexed because they are not informative [Ver04]. It is a list with all the words of tools, connection and articulation (pronouns, articles, conjunctions, prepositions, etc.). (Example: في ، ان ،على ، التي ، عن ، الذي ، مع، فى ،بعد ، بين ، هذه ، هذا ، انه ، منذ ، ما ، لم..).

Light stemming. Using words as linguistic unity is possible, but also raises a number of problems of ambiguity in the morphological analysis, the fact that Arabic (unlike the Latin languages) is an inflected language, and strongly differentiable agglutinative, articles, prepositions and pronouns stick to adjectives, nouns, verbs. To resolve the ambiguity [Bou05] showed that stemming is a very useful preprocessing, which involves finding the root of each word. It makes a deletion of prefix and suffix to identify the root word. These suffixes and prefixes are grouped in a dictionary. Since most of the Arabic words have a root with three or four letters, keeping the word at least three letters will allow us to preserve the integrity of sense. So we conducted light stemming by identifying prefixes and suffixes that were added to the word. We use the list of prefixes and suffixes proposed by [Dar03], it was determined by a frequency calculation on a corpus of Arabic articles. This list includes prefixes and suffixes commonly used in the Arabic language such as conjunctions, verbal prefixes, possessive pronouns, pronouns name or verbal suffixes expressing the plural and so on.

Table 1. Prefixes and suffixes list.

Prefixes							
والـ	بتـ	وتـ	بمـ	كمـ	للـ	فيـ	لا
فالـ	يتـ	ستـ	لمـ	فمـ	ليـ	وا	با
بالـ	متـ	نتـ	ومـ	الـ	ويـ	فا	
Suffixes							
ـات	وه	ته	هم	نا	ين	ـه	ا
وا	ـان	تم	هن	تك	يه	ـي	ون
تي	كم	ها					

2.2 Automatic extraction of "candidate terms"

After preparing the corpus, we move to the extraction step of ontology elements. The processing is done in two passages. In the first; we will extract all the terms (one or more words) used to denote concepts in the domain, using the method of "repeated segments" based on the following prepositions: *A significant term is used several times in a specialized text.*

- Terms can be complex, that are composed of several words used individually (ex. جملة اسمية) .
- Complex terms are constructed using a finite number of sequences of words.

In the second passage; we will seek the pairs of terms that co-occur more frequently in the corpus. The result of this processing provides us with a list of pairs of terms that will be used to update the ontology. Therefore, the objective of the first pass is to identify the terms that denote the concepts related to the domain, however the second pass is to identify among these terms, couples who have links with elements of the ontology.

Applying the method of "repeated segments". It is a statistical technique for extracting information from texts unlabelled. The repetition of these segments indicates that these can be used to denote concepts of domain of the corpus. A text segment consists of one or more words and delimiters are punctuation marks or spaces. The method performs an index of all words in the text by assigning a code corresponding to their positions in the corpus. Then it identifies of all repeated segments in a window of four words (number of four is chosen on the principle that a term denoting a concept contains a maximum of four words) in limiting itself to the same sentence. During this phase, redundancies are eliminated by removing the segments included in others with the same number of occurrences. At this step a large number of segments are extracted, some of which are incorrect. All of these segments are then filtered to remove unwanted segments and retain only those who are selected as candidate terms. In our approach, we use two filters; filter of weights [Her06] and a cutting filter[3]. The weighting filter is used to select terms with enough weight with respect to this weighting; it is a global threshold and fixed indicating the relevance (a relevant term is used several times in a specialized text). The weight is measured by the total frequency of a term; it is the total number of occurrences of the word in the corpus. If this *frequency* exceeds a *global threshold*, then the term is part of the domain.

The "cut filter" removes the segments containing certain words such as verbs, named entities, numbers into letters or other. The words of "cut filter" may be present at the beginning, the end and within the segment. The list of words of the filter can be easily adapted and expanded by the user depending on the specifics of the corpus treated. The words of the "cut filter" cannot be present in a segment after application of this filter.

[3] Used in the MANTEX (it is a system of terminology extraction from texts unlabelled. [RoF02]

Applying the method of "co-occurrence". The technique is based on the extraction of binary cooccurrents or pairs of terms that meet one of the other more frequently than by chance and these two terms were included in the list found in the previous phase (phase detection of repeated segments). The method starts by identifying cooccurrents of a given term in a window of fixed size (example ten words) and in the same sentence, examining the cooccurrents relative to the target term. The method measures the attraction in pairs (the terms in some order) and not in pairs. Pair {جملة, اسم} corresponds to two pairs <اسم , جملة > (جملة is the first term and اسم appears to the left in the text) et <جملة , اسم > (This time it is جملة than appears in the left).
Finally, we will select the cooccurrents with a frequency exceeding a statistically significant frequency due to chance. A numerical threshold of 80%[4] is defined a priori to estimate a relation between two terms is significant.

2.3 Update of the ontology

The principle of the approach is to compare the pair of candidate terms extracted (<t1,t2>) with the labels of the ontology concepts, we find four possible cases; t1 (t2) belongs to the labels of ontology and t2 (t1) is not, t1 and t2 are in the same time labels of the ontology, t1 and t2 or not belong to the labels of the ontology.

Relation by linguistic marker. To identify relations between terms, we will study the context surrounding these terms in a small window (eg, four words) [Koo03]. From this context the method will look for lexico-syntactic elements for identifying a relation between them. These elements are called linguistic markers[5].
Example « T1 is-a T2 », « T1 part-of T2 » ,...

But as the same relation can be expressed by different markers so they are organized into categories or separate lists depending on the type of relation to be extracted, which will be incremented progressively.

Thus we have in each list (or category), a kind of paradigm of linguistic units which are sometimes heterogeneous categories (nouns, verbs, function words or grammatical, etc.). But always it fulfills the same functions for the relation type.

- Hyponymy or Generalization relation « is-a » : list = {**هو، هي ، هم، ...**}
- Meronymy relation «part-of » : list= { **تتألف-من ، تنقسم-الى ، تتكون-من،**}

Accordingly to the specific morphology of Arabic at the vocalization and agglutination, the list of markers should be clustered all forms and other morphological variants likely to be encountered in the texts. We can add new relations and to update the lists of pre-existing relations. The process of updating the ontology is as follows:

- If one term of the pair is found among the labels of the ontology concepts, the second term of the pair will be proposed for a new concept in the ontology and will be linked to the first concept for a relation defined by linguistic marker.

[4] The numerical threshold used in the "Xtract" extractor is 80%. [Sma93]
[5] CAMELEON is a software research of lexical relations from linguistic markers. [Ség01]

- If both terms are among the labels of the ontology concepts and there was no relation between these two concepts, a new relation will be proposed from marker linguistic.
- In case where neither the first nor the second term do not belong to the ontology labels. The process does nothing and let these cases for future running.

Hierarchical relation. If the linguistic markers are absent in the context of words, the approach based on a parent-child relation where the parent term is more general than the child term. This relation between terms is extracted from the asymmetric co-occurrence of terms. The relation is characterized by the following two rules: $P(x/y) \geq 0.8$ and $P(y/x) < P(x/y)$; P(x/y) is the probability of term 'x' occurrence then the term 'y', inversely for P(y/x) [HeM06]. First rule ensures that both terms appear together enough (ie 80% of cases). According to the second rule, x subsumes y where the probability of occurrence of x before y is upper than the reverse. Using the transitive property of the relation we can eliminate some relations, e.g. if the relation "a" subsumes "b", "a" subsumes "c" and "b" subsumes "c" are extracted, the relation "a" subsumes "c" can be deleted because it is deductible from the other two [Her06]. However, the process of updating the ontology is as follows:

- If the first term (or second) is found among the labels of the ontology concepts and the second (or first) term of the couple is not, then it will be proposed a new *son-concept* (*father-concept)* related to the first (second) concept by subsumption relation “is-a”.
- In the case where both terms are among the labels of the ontology concepts and there was no relation between these two concepts, a new relation of subsumption “is-a” will be proposed.
- In case where neither the first nor the second term do not belong to the ontology labels. The process does nothing and let these cases for future running.

3. Experimentation and results

We were able to test the approach using the Python programming language, due to its power and through its NLTK[6] (Natural Language Toolkit) library.

3.1 Constitution of corpus

We selected a sample of texts from documents written in Arabic sought in the following resources: books on Arabic linguistics, and journal articles (N°7 and N°8 of AL-LISANIYYAT) published by the CRSTDLA[7] in Arabic language and through

[6] http://nltk.sourceforge.net/index.php/Main_Page
[7] Center for Scientific Research and Technical Development of Arabic Language (Algiers)

the Web by introducing specific keywords related to the domain in the search engine "Google". The queries used are:

نظرية اللسانية ،علم الدلالـة، الألفاظ في النحو العربي، اللسانيات الحديثة، قواعد في اللغة العربية، الأبنية ودورها في اللغة العربية ، الأوزان ، خصائص اللغة العربية، النحو العربي.

The documents found are downloaded, selected and prepared manually (by deleting tables, diagrams and graphs), these documents are usually texts compiled in Word or PDF. We must transform them into simpler format "plain text" (.txt). The following table 2 shows our corpus characteristics.

Table 2. Technique characteristics of corpus.

Total number of documents	57
Total number of words	468 554
Total size (K byte)	2 742 Kb

3.2 Preprocessing

Segmentation and Normalization. We have segmented the texts to word sequences by detecting word delimiters such as spaces or punctuation. We also used the list of Arabic punctuation symbols, as the following: ["،","."," ؟","!","...","..",":","؛"]. In the normalization, we removed all the elements that do not provide information and increase the processing time, such as special characters, numbers, non-Arabic words, abbreviations, single letters and deletion of vowels. Example of special characters: ["–","/",",","«","+","%",…].

Result. 417 059 words are selected and 51 495 words are deleted (11%).

Deletion of stop words (1). We have made this list of stop words from the corpus on two principles: their frequency and their information content. We have sorted the most used words in the corpus according to their frequency, and then we manually selected among them the words that do not have information related to the domain. (In total we sorted 455 stopwords).

Result. The list is not exhaustive, so we always update it with new words or new morphologicals forms of the same word. The result of processing (repeated segments) is strongly dependent on this step. We have eliminated 116 137 words (27.9%).

Light Stemming. We have removed the prefixes and suffixes following predefined list, (Table 1). This list is stored in two files (prefixes and suffixes file).

Result. In the results, we found instances where the same word appears again in several morphological forms which will decrease the performance of the processing.

Suggestion. To remedy this problem, we can use a tool for morphological analysis in this step to complete the lemmatization which will significantly improve the quality of processing.

Deletion of stop words (2). We need to eliminate stop words again, since in the results of light stemming we found these words again after deleting some of the prefixes and suffixes: Example (following cases are present: بعده-بعد ، اخرى-الاخرى)

Result. 261 715 words are found and 39 207 words are removed (13%).

3.3 Processing

Extraction of "repeated-segments". We set the following parameters:

- Segment size = 4 words. It indicates the maximum size of a complex term, usually a complex term in Arabic is made up of 4 words.
- Weighting threshold: The weight of a term is calculated by the total frequency, is the total number of occurrences in the corpus. Threshold weight of a simple word is = 100. Threshold weight of a compound term is = 20. The number 100 and 20 are randomly selected relatively to the corpus size.

Result. The program extracts 281 200 different segments, but it only selects a list of 445 segments in accordance with the thresholds defined above. In analyzing this list, we have identified the following comments:

1. Words appear that are outside domain (personal names, object names ...). We can update the list of stop words by these words and to redo processing.
2. Two morphological forms of same word are identified as two different segments. Example (لغه ، لغوى، لغات ، للغه) (حرف ، حروف) (عنصر ، عناصر) . We can regroup the different morphological forms in the same form then replace them in the corpus and repeat the processing.

The following table shows a sample of selected segments:

Table 3. Sample of selected segments.

Segment	Frequency	Segment	Frequency	Segment	Frequency
لغه	5071	فاعل	592	مفعول مطلق	84
فعل	2449	ظاهر	579	جمل اسم	83
اسم	1938	ضمير	575	علام رفع ضمه	78
…	…	…	…	…	…

Extraction of "co-occurrents". We set the following parameters:

- Window size of co-occurrence = 10 words.
- Co-occurrence threshold = 80% (percentage of appearance two terms together).
- Co-frequency threshold = 100 (number of appearance two terms together).

The program gives the result in a marked file where each line contains the co-occurring, their frequency and their co-frequency. As the following example:

< t1="نصب" t2="فتح" Ft1="672" Ft2="129" CF="211"/>
< t1="اسم" t2="فعل" Ft1="1938" Ft2="2449" CF="210"/>

Suggestion. This result file must be validated by an expert (a linguist).

4. Conclusion

In this paper, we have shown an approach for the automatic construction of ontology from a corpus of domain "Arabic linguistics". We reused information extraction techniques for extracting new terms that will denote elements of the ontology (concept, relation). To analyze the texts of the corpus, two statistical methods were used, the “repeated segments” to identify the candidate terms and “co-occurrence” to the updating of ontology. So, we have formed a domain corpus by the recovery of text from articles of journals and books of the domain and also the collection of documents over the Web. This corpus was preprocessed to remove some ambiguity, reduce the number of transactions and adapt the corpus according to our aim.

Many perspectives are offered based on our work, among them; we proposed an ontology that represents the fundamentals notions of Arabic linguistics, this ontology can be useful for developing NLP tools that analyze Arabic texts. A second perspective would be to use our techniques and statistical methods for information extraction on Arabic texts for other works (e.g. terminology extraction, creation of electronic dictionaries and thesaurus ...).

References

[AbD08] Ramzi Abbès, Joseph Dichy « Extraction automatique de fréquences lexicales en arabe » JADT 2008 :« 9eme Journées internationales d’Analyse statistique des Données Textuelles » Université Lumière Lyon 2, ICAR-CNRS.

[BoA03] Didier Bourigault et Nathalie Aussenac-Gilles. «Construction d’ontologies à partir de textes ». conférence sur le traitement automatique des langues (TALN), France, Juin 2003.

[Dar03] Darwish K « *Probabilistic methods for searching OCR-Degraded Arabic Text* » Thèse de Doctorat Université de Maryland 2003.

[Dou05]. F. S. Douzidia, G. Lapalme « *Un système de résumé de texte en arabe* » université de Montréal exposé en deuxième conférence International de "l'Ingénierie de la Langue et Ingénierie de l'Arabe " Alger 2005.

[Far03] : Farrar, William D. Lewis, and D. Terence « *An Ontology for Linguistic Annotation* **»** Department of Linguistics, University of Arizona 2003.

[HeM06] N. Hernandez, J. Mothe « **TtoO**: une méthodologie de construction d’ontologie de domaine à partir d’un thésaurus et d’un corpus de référence » IRIT, Toulouse, 2006.

[Her06] Nathalie HERNANDEZ « Ontologies de domaine pour la modélisation du contexte en recherche d’information » Thèse de Doctorat à l’Université Paul Sabatier France 2006.

[Koo03] S. Koo, S.Y. Lim, S.J. Lee, « Building an Ontology based on Hub Words for Informational Retrieval », the IEEE/WIC International Conference on Web Intelligence, 2003.

[Mar03] Elizabeth Marshman «Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie » Janvier 2003, "Observatoire de linguistique Sens-Texte" (OLST) de l'Université de Montréal.

[RoF02] F. Rousselot et P. Frath, « Terminologie et Intelligence Artificielle » (12èmes rencontres linguistiques), Presses Universitaires de Caen, 2002.

[Ség01] Patrick Séguéla « Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques » thèse TOULOUSE III. 2001.

[Sma93] Frank. Smadja, « Retrieving collocations from text: Xtract, Computational Linguistics », université de Columbia 1993.

[Ver04] Jacques Vergne « Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource » JADT 2004 :« 7eme Journées internationales d'Analyse statistique des Données Textuelles » GREYC – Université de Caen.