Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 73 (2015) 7 – 15

Procedia Computer Science

The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Building ontology from texts

Bedr-eddine Benaissa[a,*] , Djelloul Bouchiha[a], Amine Zouaoui[b], Noureddine Doumi[d]

[a,b] *Dept. Mathematics and Computer Science, Ctr Univ Naama, Algeria*
[c] *Dept. Mathematics and Computer Science, University of Sidi Bel-Abbes, Algeria*
[d] *University Dr. Tahar Moulay of Saida, Algeria*

Abstract

The purpose of this paper is to present an approach to create semi-automatically ontology from Arabic texts. The whole process is supervised by a linguistic expert. Our involvement in this project focused on a lexical ontology, taking as model the WordNet ontology, and as input source, the "Arabic verbs" of a contemporary monolingual dictionary (معجم الغني) /mʿjm Alγny/ in the form a lexical database. The verb, pivot of a sentence, is our goal in creating concepts, by adopting the synset as our meaning representation model. The Markov clustering algorithm of a graph, generated by the defining verbs, obtained from the transitive closure, allowed us to detect similar verbs and to identify as well, for a given verbal entry, all of its synonyms. A tool has been implemented, and experiments have been carried out to evaluate and show efficiency of the proposed approach.

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Keywords: Ontology Learning, (Lexical) Ontology, Arabic Texts, NLP, Markov Clustering Algorithm, WordNet.

1. Introduction

Many researchers are investigating in the study of natural language, where they have to confront hard issues. This difficulty is due to the insufficiency of the conventional representation models (logic, production rules, semantic

* Corresponding author. Tel.: +213-(0)55-430-4141.
E-mail address: b_benaissa@mail.univ-tlemcen.dz

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)
doi:10.1016/j.procs.2015.12.042

networks ...) to correctly represent linguistic concepts. Seeing the increasing number of digital textual documents, researchers were trying to find technical solutions to control this mass of information. The appearance of metadata is the result of their studies. This new approach motivated the revelation of structured languages such as SGML, XML, RDF. This new development contributed to the construction of ontologies.

The construction of ontologies from texts is a subdomain the ontology learning field. In the context of Semantic Web, ontologies are used primarily for the semantic annotation of resources and for structuring knowledge bases.

This issue is indeed an important new challenge for both Natural Language Processing (NLP) and for Knowledge Engineering (KE). Several works already exist in this area for different languages such as English and French. Unfortunately, for the Arabic language, works are still in the beginning. So we try to define a methodology for building ontology for the Arabic language.

In this paper we propose an approach to construct the ontology concepts based on synonymy relation between the verbs of the Arabic language. Thus, we define the basic principles necessary for understanding our approach and then we present experiments and results, we conclude with an evaluation of the adopted method, and we discuss its usefulness to various fields.

The ultimate goal of this work is to present different theoretical and practical "basic" interests to create ontology, specifically a lexical ontology with a perspective to be used in NLP system.

The remainder of this paper is organized as follows: the next section presents a survey of the major recent works in the field of ontology learning. Section 3 explains the notions in relation with our problem. Section 4 describes the proposed approach. Section 5 describes the implemented tool. An experimental study showing the effectiveness of our approach is presented in Section 6. Section 7 concludes our work.

2. Related work

Texts are rich in knowledge and build up a shared vocabulary between a large community of a domain. Our issue is to acquire, from a text, a set of useful knowledge to build ontology. This belongs to the "ontology learning" researches. Several recent contributions were the subjects of papers in the ontology learning field.

Kamel et al.[11] propose an approach based on Web data sources, including forms that are a source of structured data. Their study uses various properties of these documents, with the combination of a layout analysis, a linguistic analysis and semantic annotation. They propose to construct a domain ontology in two stages: the first is to build a core ontology and the second is to enrich it.

Silva et al.[13] propose an alignment in several stages: in the first stage, they gather the terms of the first three levels of the domain ontology, and associate them to the concepts of the basic used ontology. Then other preliminary steps are also considered, such as extraction and cleaning of fragments. The alignment is then applied with selected measures, based on the OMN standard (Naive Ontology Mapping) used by the FOAM tool[6].

Carvalho et al.[3] look for implicit information in the domain ontology, and operate the way it can be extracted by improving various processes, notably the alignment. This approach uses data mining techniques to extract new terms and relations from ontologies, to allow their semantic improvement, by enriching ontologies with these elements.

Carvalho et al.[4] consider the enrichment of ontologies with relations and implicit terms contained in the definitions of ontologies, as well as the association of the ontology concepts to the categories of the basic ontologies.

Some old ontology learning systems are described in[12].

3. Background

Lexical ontology: Lexical ontologies can be considered, as well, as a lexicon or as an ontology[10], and are significantly different from conventional ontology[10]. They are not based on a specific domain, but they are intended to provide structured knowledge about lexical issues (words) of a language by linking them to their meanings[15].

Arabic WordNet (AWN): The Arabic WordNet is a lexical database. Its design is based on WordNet Princeton and it is built using methods developed for EuroWordNet and connected with the ontology SUMO (Suggested Upper Merged Ontology)[2]. Most Synsets AWN must match their counterparts in the English WordNet and the entire topology of the two wordnets should be similar.

Markov Cluster aLgorithm (MCL) & Clustering: Given an adjacency matrix resulting from a synonymy graph. MCL[14] finds groups by simulating random paths in a graph by counting alternatively the random paths of the greater length, and by increasing the probability of intra-cluster paths. MCL can be described briefly in five steps:

1. Take the adjacency matrix A of the graph,
2. Normalizing each column A to 1 to obtain a stochastic matrix S,
3. Calculate S2,
4. Take "e" power of each element of S2 and then normalize each column to 1,
5. Return to (2) until the MCL converges to an idempotent matrix - steps (2) and (3).

4. Proposed approach

Our ultimate goal is to create a lexical ontology for the Arabic language. The question that arises is, why create such ontology, while AWN already exists?

4.1. Answer

The creation of a wordnet, and the creation of most ontologies, is typically manual and involves a lot of human effort. Some authors[5] propose the translation of the Princeton WordNet to wordnets in other languages (eg. Arabic WordNet[2]. However, if it could be suitable for several applications, a problem arises because different languages represent different socio-cultural backgrounds, which do not perceive exactly the same part of the lexicon, and even if they appear to be common, several concepts are lexicalized differently[11]. For example, in AWN, the verb "حَلَّلَ" does not find the term "زَوَّجَ" in his synset.

4.2. Goal

The construction of a lexical ontology from Arabic texts should present the concepts of modern Arabic language and various semantic relations between these concepts. However, it is limited in this paper to discover the lexical concepts or in other words, Arabic synsets.

4.3. The construction process

Fig.1. shows an overview of the building process of the target lexical ontology.

Fig. 1. An overview of the lexical ontology construction process

4.4. Solution

We propose a semi-automatic approach to build a lexical ontology, based on (1) a representation graph (hierarchy) of synonyms found in the dictionary, and (2) a clustering method to exploit the graph, and which will allow the automatic construction of groups of synonyms. Each constructed group will be considered a synset.

In its current version, our approach focuses on verbs.

4.5. Hypothesis

As the approach adopted by[8], we recognize that "the grouping in clusters from the synonyms graph of a dictionary provides groups of fairly similar synonyms that can be considered synsets".

We add to this hypothesis the conservation of hierarchical relations (subsumption) between the detected synsets.

4.6. Basic Corpus

For our study, we took an Arabic dictionary (معجم الغني)*. In a first perspective, we focused on Arabic verbs. The proposed solution will allow us to extract synonymy relations (basic relation for the construction of concepts) by morphologically analyzing of an Arabic verb in input.

Fig. 2 shows the definition of the verb « اِمْتَحَنَ» /Aim.taHana/ in the dictionary (معجم الغني) /mʤjm Alγny/.

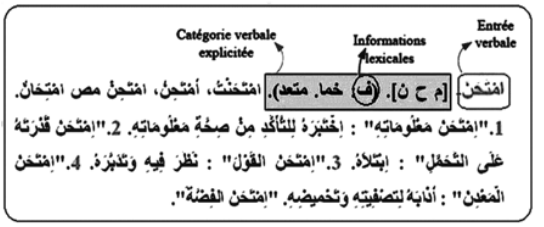

Fig. 2. Definition of the verb « اِمْتَحَنَ» /Aim.taHana/ in the dictionary (معجم الغني)

4.7. Algorithm

Our algorithm looks for direct synonyms of a verbal entry of the dictionary « معجم الغني » /mʤjm Alγny/. Then the defining verbs become inputs to their towers, and we continue the search for each one of the defining verbs, and so on ... taking only one time the extracted verb so that we can stop.

Construction of the graph: the construction of the graph (hierarchy of terms) is done through the exploitation of the natural structure Defined-Verb $\Leftrightarrow$ Defining-Verb derived from the dictionary in order to derive the most of synonyms corresponding to an entry.

Note that: The extraction of the defining verbs is done using the defined patterns dedicated to the selected dictionary. Thus, considering VE a verbal entry of the dictionary, the first lexical pattern (or morphological) was defined:

$$[[[[\ \ldots.\ \text{كلمة}_4\ ]،\ \text{كلمة}_3\ ]،\ \text{كلمة}_2\ ]،\ \text{كلمة}_1\ ]\ :\ V_E \qquad (1)$$

كلمة1 /klmħ1/, كلمة2/klmħ2/, … are the defining words of the verbal entry VE.

We assume also the hypothesis:

"*A defining verb is a near synonym, if it is the only verb of the definition sentence*".

Construction of synsets: the construction of synsets is done by applying the Markov clustering algorithm (MCL)[15]. This is the same algorithm used by[7] to find clusters in a synonymy network.

Construction of the ontology: the construction of the lexical ontology is done through the conservation of the hierarchical links between synsets of the original graph:

- Terms of the same cluster are abbreviated in the same synset.
- A link between two clusters remains a link between the two corresponding synsets.

---

* http://lexicons.sakhr.com/

This process allows the generation of a partial lexical ontology from a single verbal entry. To build a global ontology, we simply apply the approach to all of the verbal entries of the dictionary. As a result we obtain a set of partial ontologies which must be merged into a single lexical ontology for the verbs of the Arabic language.

4.8. Example :

Given the verbal entry « اِمْتَحَنَ» /Aim.taHana/. Its definitions, according to the dictionary (معجم الغني) /mcjm Alγny/, are represented in the figure below:

a

	Contexte	Définition
1	اِمْتَحَنَ مَعْلُومَاتِهِ	اِخْتَبَرَهُ لِلتَّأَكُّدِ مِنْ صِحَّةِ مَعْلُومَاتِهِ
2	اِمْتَحَنَ قُدْرَتُهُ عَلَى التَّحَمُّلِ	اِبْتَلاَهُ
3	اِمْتَحَنَ القَوْلَ	نَظَرَ فِيهِ وَتَدَبَّرَهُ
4	اِمْتَحَنَ المُعَلِّمُ	أَدَابَهُ لِتَصْفِيَتِهِ وَتَحْصِيلِهِ

b

Contexte	Définition
اِبْتَلاَهُ عَنْ قُرْبٍ	اِخْتَبَرَهُ
اِبتَلَى أُمُورَ الحَيَاةِ	عَرَفَهَا
اِبْتَلاَهُ بِعِلَّةٍ	اِخْتَبَرَهُ، اِمْتَحَنَهُ

Fig. 3. (a) The defining verbs of the verb « اِمْتَحَنَ»; (b) The defining verbs of the verb « اِبْتَلَى»

We note that here, there is only one defining verb according to our patterns «اِبْتَلَى» /Aib.talaγa/. For this verb, looking for its defining verbs, we obtain the extracted verbs : « اِخْتَبَرَ» /Aix.tabara/, « عَرَفَ» /çarafa/. We continue to look for synonyms of the two verbs previously found.

The first verb « اِخْتَبَرَ» /Aix.tabara/ has two defining verbs according to our patterns « اِمْتَحَنَ» /Aim.taHana/, «جَرَّبَ » /jar~aba/ (see Fig.4):

a

Contexte	Définition
اِخْتَبَرَ ذَكَاءهُ	اِمْتَحَنَهُ، جَرَّبَهُ
اِخْتَبَرَ حَقِيقَةَ الأَمْرِ	عَلِمَهُ عَلَى حَقِيقَتِهِ
اِخْتَبَرَ الرَّجُلُ لأَهْلِهِ	اِشْتَرَى لَهُمْ إِدَاماً أَوْ لَحْماً

b

Contexte	Définition
عَرَفَ الشَّيْءَ	أَدْرَكَهُ بِعِلْمِهِ، خَبَّرَهُ
عَرَفَ بِذَنْبِهِ	اِعْتَرَفَ بِهِ
عَرَفَ لِلأَمْرِ	صَبَرَ
لأُعَرِّفَنَّكَ عَلَيْهِ	لأُجَازِيَنَّكَ عَلَيْهِ

Fig. 4. (a) The defining verbs of the verb « اِخْتَبَرَ»; (b) The defining verbs of the verb « عَرَفَ»

The second verb « عَرَفَ» /çarafa/, possesses « خَبَّرَ» /xab~ara/ and « صَبَرَ» /Sabara/ (see Fig. 4): and so on, until exhausting the definitions of each verb found. Thereby, we obtain a directed sub-graph in which vertex is the verbal entry and arcs represent synonymy relation. That's part of the sub-graph generated by the verbal entry « اِمْتَحَنَ» .

"In the following, we note that the global synonymy graph Gs is a sub-graph of G_D. In Gs we keep only the synonymy relation between nodes. Let us mark also, that in the global synonymy graph Gs (derived from the G_D graph by applying patterns), there is a number of synonymy sub-graphs"

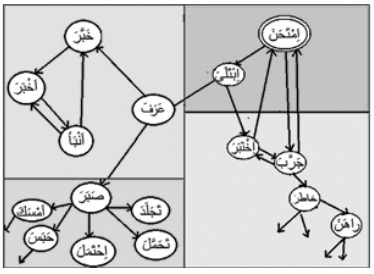

Fig. 5. Graph GS of the verb « اِمْتَحَنَ»/Aim.taHana/

5. Implementation

To implement our approach, we have developed a tool called OntoArab-Maker. Fig.6. shows the various processing steps and the interaction of OntoArab-Maker's resources.

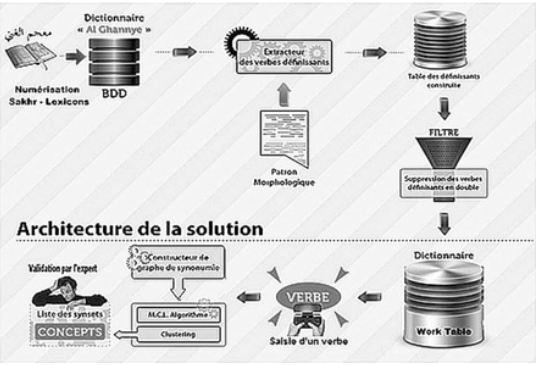

Fig. 6. Architecture of OntoArab-Maker

The Extractor uses the syntactic patterns presented above to extract the defining verbs from the input dictionary. This step is immediately followed by a cleaning operation which involves the removal of the duplicate defining verbs. Other procedures are applied to clean and prepare the table to the treatment phase. The recovered database contains the verbs table without double.

6. Results and evaluation

We present in this section they obtained results on a sample of selected verbs. Note that the proposed approach, although wholly automatic, requires a validation step of the calculated synsets by an expert of the Arabic language. So we thought to integrate him into a semi-automatic solution for the construction of synsets.

We opted for a fully manual evaluation of synsets of the selected sample. The expert is responsible for evaluating, according to its estimation, the accuracy of each synset and the eventual anomalies.

Our evaluation was focused on a sample of two verbs: « حَلَّلَ » /Hal~ala/ and « اِمْتَحَنَ » /Aim.taHana/. The evaluation process was based on two criteria: number of meanings conveyed in a calculated synset and the number of impertinent verbs in the synset.

A synset is a set of near synonyms carrying a unique meaning (sense). A calculated synset (cluster) is considered correct, if it conveys only one sense. If the number of senses of a sysnset is greater than 1, then the clustering is less accurate.

The expert, when evaluating the calculated synset, divides it into two sub-synsets or more, if he judges that the number of conveyed meaning is greater than 1.

The impertinent verbs appointed by the expert are verbs that should not be included in a calculated synset and can not belong to any sub-synset proposed by the expert.

The evaluation results were reported in Table 1 and 2 as follows:

- Column 2: Synset (cluster), calculated by our system. Each cell contains the words (verbs) corresponding to a synset calculated by the system. In the same cell, these verbs are organized by the expert in one or more sub-synsets. Terms, qualified by the expert as impertinent (does not belong to the synset), are strikethrough.
- Column 3: Number of senses evoked by the calculated synset. This number is determined of course by the expert.
- Column 4: Number of impertinent verbs in the synset, even if this synset is fragmented into several other correct sub-synsets.
- Column 5: Expert appraisal for a synset, based on the values in column 3 and column 4.

As we have noted that there is an inverse proportional relationship between the expert appraisal and the values in columns 3 and 4, we proposed a formula to express the success rate (appreciation of the expert) for identifying a synset: T = (((1/ C3) + (1/(C4 + 1))) / 2)* 100% ; with C3 and C4 are the values in columns 3 and 4 respectively.

For the test protocol, we started using the usual values of the original approach[8], with r = 1.6 (inflation parameter) and e = 2 (expansion parameter).

The following two tables show the results obtained with these parameters and the expert evaluation:

Table 1: Evaluation on the verb « حَلَّل » /Hal~ala/ with r = 1.6

N°	Synset calculé à partir d'une entrée verbale	Différents sens évoqués du Synset	Nbr. de verbes non pertinents	Taux de réussite d'identification d'un synset
01	حَلَّلَ، زَوَّجَ، شَرَحَ، فَسَّرَ، بَيَّنَ، شَرَّحَ، أبانَ	1. Analyser et expliquer	0	100%
02	فَكَّ، حَلَّ، فَصَلَ، عَدَّلَ، سَوَّىَ، أَشْرَعَ، اِسْتَبْدَلَ، غَيَّرَ، حَوَّلَ، نَقَلَ، أحالَ، أَفْشَىَ أَظْهَرَ، أَوْضَحَ، أَبْرَزَ، أَخْرَجَ، أَصْدَرَ، أذاعَ، أَحْكَمَ، نَثَرَ، اِتَّضَحَ، بانَ، ظَهَرَ، أَرْشَدَ، نَشَرَ، أقامَ، أَنْضَجَ، زَكَّىَ ، قَبِلَ، سَحَبَ، أَوْرَقَ	1. Dissocier pour réparer 2. Echanger 3. Faire apparaître 4. Eclaircir	3	25%
03	سَدَّدَ، صَوَّبَ، فَتَحَ، وَجَّهَ، أَصْلَحَ، عَلَّمَ شَرَعَ، قَدَّمَ، اِتَّخَذَ،	1. Viser 2. Entreprendre	0	75%
04	فَضَّ، فَرَّقَ، شَتَّتَ، قَطَّعَ، قَسَّمَ، وَزَّعَ، سَكَبَ، ذَرَفَ، أَفاضَ، صَبَّ، سالَ، أَسالَ، اِنْحَدَرَ، سَرَّحَ حَسَّنَ، زَيَّنَ ، رَقَّ، خَوَّنَ، اِشْتاقَ	1. Disperser 2. Verser 3. Embellir	2	33.33%
05	نالَ، اِسْتَغْرَقَ،أَنْهَىَ، قَتَلَ، وَفَىَ، أَهْلَكَ، قَضَىَ، أَفْنَىَ، أَدَّىَ، أَنْفَدَ، أَبْلَغَ، أَوْصَىَ، أَخَذَ ، لَمَرَ، أرادَ،	1. Achever 2. Vouloir	1	50%
06	رَطَّبَ، بَلَّ، أَرْطَبَ، ألانَ، نَعَّمَ	1. Humidifier	0	100%
07	عَيَّنَ، جَعَلَ، اِعْتَمَدَ، خَلَقَ، صَنَعَ، وَضَعَ، أَلْقَىَ، قَصَدَ رَمَىَ،	1. Faire en sorte que …	0	100%
08	كَسَّرَ، حَطَمَ، هَشَّمَ، كَسَرَ، فَتَّتَ، دَقَّ، حانَ	1. Casser et Broyer	1	75%
09	أَتَمَّ، اِمْتَلَأَ، اِكْتَمَلَ	1. Etre rempli …	0	100%
10	حَدَثَ، حَصَلَ، اِقْتَحَمَ بَدَأَ، اِنْطَلَقَ، أَنْشَأَ، اِنْتَقَلَ،	1. Commencer	1	75%
11	خاضَ، أَسْرَعَ، خَلَطَ، حَرَّكَ	1. Plonger dans …	0	100%
12	مَهَّدَ، مَهَدَ، بَسَطَ، سَهَّلَ	1. Introduire	0	100%

Table 2: Evaluation on the verb « اِمْتَحَنَ » /Aim.taHana/ with r = 1.6

N°	Synset calculé à partir d'une entrée verbale	Différents sens évoqués du Synset	verbes non pertinents	Taux de réussite d'identification d'un synset
01	{اِمْتَحَنَ،اِبْتَلَىَ،اِخْتَبَرَ،راهَنَ،عَرَفَ} {خاطَرَ، جَرَّبَ،خَبَّرَ } {أَخْبَرَ، أَنْبَأَ}	1 - Examiner, 2 - s'expérimenter 3 - informer	0	66.67%
02	{صَبَرَ، تَجَلَّدَ، تَحَمَّلَ، اِحْتَمَلَ، حَبَسَ، اَمْسَكَ} { تَصَبَّرَ، أَطاقَ، اَغْضَىَ، سَكَتَ} {اِرْتَحَلَ، شَكَرَ، قَبَضَ، كَفَّ، اِمْتَنَعَ، صَمَتَ } أَظْلَمَ	1 - Patienter 2 - se faire tout petit 3- tourner le dos à	1	41.67%
03	ضَبَطَ، أَتْقَنَ، أَحْكَمَ، كَبَحَ}	1 - maîtriser	0	100%
04	قَبِلَ، اتَىَ، حَلَّ، هَبَّ، وَقَعَ	Venir	1	75%
05	صَفَحَ، قَلَبَ، تَصَفَحَ، رَدَّ، نَظَرَ	1 - pardonner	0	100%

Statistics on the results obtained with r = 1.6 for the verbs « حَلَّلَ » /Hal~ala/ and « اِمْتَحَنَ » /Aim.taHana/

- Average of senses evoked by synset: 1.64.
- Percentage of synsets with only one sense: 58%.
- Percentage of synsets with two senses: 11.76%.

- Percentage of synsets having more than two senses: 23%.
- Average number of impertinent verbs in synset: 0.58.
- Synsets rate with 0 impertinent verbs: 58%.
- Synsets rate with a single impertinent verb: 29.41%.
- Synsets rate with two or more impertinent verbs: 11.76%.

With this sample, we note that the majority of calculated synsets has only one sense and has at most a single impertinent verb. We therefore consider these results satisfactory, since a number of impertinent verbs, less than two (zero or one), remain very acceptable.

To determine the optimal inflation value ***r***, i.e. where our algorithm performs better, we decided to launch the clustering process with different values of the parameter ***r*** to obtain a finer granularity. One question remains raised: which measure should we use to evaluate our algorithm?

The synsets identification issue is similar to the information retrieval one[1], where the evaluation of the retrieving process is performed using two metrics: Precision and Recall*.

Thus, we define in our approach the Recall as the proportion of the relevant synsets found by the system for a given value of the parameter ***r***, of all the relevant synsets offered by the linguistic expert (=8 for the verb « اِمْتَحَنَ » `/Aim.taHana/`). Whereas the precision is the proportion of the relevant synsets found by the system for a given value of ***r***, of all synsets proposed by the linguistic expert (=10 for the verb « اِمْتَحَنَ » `/Aim.taHana/`).

The F-measure combines the precision and recall, and is calculated as follows:

$$\text{F-measure} = (2 * \text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}). \quad (2)$$

The following graph describes the evolution of Recall, Precision and F-measure of the synsets identification process of the verbal entry « اِمْتَحَنَ » `/Aim.taHana/` with different values of ***r***.

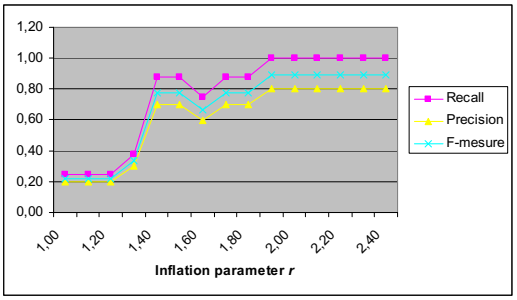

Fig. 7. The values Recall, Precision and F-measure with different inflation values for the verb « اِمْتَحَنَ » /Aim.taHana/

For ***r*** within the interval 0 - 0.9, no synset was identified.

We also note that for ***r*** greater than or equal to 1.9, the F-measure becomes maximum. This means that in terms of number of returned synsets, the results are more accurate with *r>=1.9*.

Therefore, with *r >= 1.9*, the clustering process suggests a much finer granularity. However, we note that even if the granularity is more appropriate with inflation equal to 1.9, some verbs are in the wrong place (in the wrong synset). In other words, we must exchange one or more verb between two synsets to get totally correct synsets. We believe that some discovered clusters by the MCL algorithm, are correct from a mathematical point of view, but they are sometimes semantically incorrect. However, we must remember that in a semi-automatic operating environment, automatically discovered clusters are a great help, and would effectively support the expert in his task of creating concepts of a lexical ontology.

The obtained results show that the proposed synsets (clusters) remain perfectly usable in a semi-automatic

* http://en.wikipedia.org/wiki/Precision_and_recall

synsets construction solution, which requires the intervention of the expert for the validation stage.

7. Conclusions & perspectives

With the ontology, Semantic Web is able to take into account the meaning of words rather than their syntaxes. The fully manual construction of ontology is a difficult task, complicated and requires a lot of time and resources. The use of automatic or semi-automatic methods has become indispensable. However, the use of experts for the validation of the results allows to achieve a more perfect and precise ontology.

Princeton WordNet is our ontology model for the representation of concepts and their relationships. Automatic discovery of synsets was our goal in this work. For this, we used a monolingual dictionary of the Arabic language to extract a set of synonyms on which we applied the MCL algorithm (Markov Clustering algorithm) for the detection of synsets.

The results were very encouraging. Thus, one can say that the clustering proved to be a good alternative to create concepts (synsets) of our lexical ontology from synonymy network of a dictionary.

It has yet to apply the approach on all verbal dictionary entries. As a result we have a set of partial ontologies which must be merged into a single lexical ontology for the verbs of Arabic language.

As future work, we plan to complete the proposed solution by the detection of relations between synsets. The use of lexical resources, other than a dictionary, is another way to improve the proposed approach. The Arabic text corpus annotated or not, allows us to do more experiments with other techniques for detecting the lexical and semantic relations.

Acknowledgement

We express our thanks to Mr. Benaissa Tedjini, Professor in Arabic linguistics at the University of Abu Bakr Belkaid – Tlemcen, ALGERIA, for the fruitful discussions about the syntax of Arabic sentences and their morphosyntactic analysis, and also to have contributed to the validation of our results.

References

1. Baeze-Yates R., and Ribeiro-Neto B.: Modern Information Retrieval. Addison-Wesley, ACM Press, Reading, MA; 1999.
2. Black W. J., Elkateb S., Fellbaum C., Alkhalifa M., Pease A., Rodríguez H., Vossen P., Introducing the Arabic WordNet project. In: 3rd Global Wordnet Conference, Jeju Island, Korea; January 2006.
3. Carvalho M. G. P., Campos L. M., Braganholo V. P., Campos M. L. M., Campos M. L. A., Extracting New Relations to Improve Ontology Reuse. Journal of Information and Data Management, Vol. (2), No. (3); 2011. p. 541–556.
4. Carvalho M. G. P., Campos M. L. M., Campos L. M., Cavalcanti M. C., OntoAlign++: a combined strategy for improving ontologies alignment. In: 6th Seminar on Ontology Research in Brazil, Belo Horizonte, Brazil; September 2013.
5. de Melo G. and Weikum G., On the utility of automatically generated wordnets. In: 4th Global WordNet Conf. (GWC), Szeged, Hungary, University of Szeged; 2008. p. 147–161.
6. Ehrig M., and Sure Y., FOAM - Framework for Ontologie Alignment and Mapping- Results of the Ontologie Alignment Evaluation Initiative. In: Workshop on Integrating Ontologies; 2005.
7. Gfeller D., Chappelier J. C. and De Los Rios P., Synonym Dictionary Improvement through Markov Clustering and Clustering Stability. In: International Symposium on Applied Stochastic Models and Data Analysis (ASMDA); 2005. p. 106–113.
8. Gonçalo O. H. and Gomes P., Towards the Automatic Creation of a Wordnet from a Term-based Lexical Network. In: 16th Workshop on Graph-based Methods for Natural Language, Uppsala, Sweden ACL; 2010. p. 10–18.
9. Gruber T., A translation Approach to Portable Ontology Specifications Knowledge Acquisition; 1993. p. 199–220.
10. Hirst G., Ontology and the lexicon. In: S. Staab and R. Studer, Handbook on Ontologies; Springer, 2004. p. 209–230.
11. Kamel M. , Aussenac-Gilles N., Buscaldi D., Comparot C, A semi-automatic approach for building ontologies from a collection of structured web documents. In: 7th International Conference on Knowledge Capture (K-CAP'13). Banff, Canada, June 2013.
12. Shamsfard M., and Barforoush A. A., The State of the Art in Ontology Learning: A Framework for Comparison. Intelligent Systems Laboratory. The Knowledge Engineering Review; 2003. Vol. 18, Issue (4).
13. Silva V. S., Campos M. L. M., Silva J. C. P., Cavalcanti M. C., An Approach for the Alignment of Biomedical Ontologies based on Foundational Ontologies. Journal of Information and Data Management; 2011, Vol. 2, N°3, p. 557–572.
14. Van Dongen, S., Graph Clustering by Flow Simulation. PhD Thesis. University of Utrecht, The Netherlands; 2000.
15. Wandmacher T., Ovchinnikova E., Krumnack U. and Dittmann H., Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In: 3rd Australasian Ontology Workshop (AOW). Gold Coast, Australia ACS; 2007, Vol. 85 of CRPIT. p. 61–69.