Using the Yago ontology as a resource for the enrichment of Named Entities in Arabic WordNet

Lahsen Abouenour[1], Karim Bouzoubaa[1], Paolo Rosso[2]
1 Mohammadia School of Engineers, Med V University
Rabat, Morocco
2 Natural Language Engineering Lab. - ELiRF, Universidad Politécnica
Valencia, Spain
E-mail: abouenour@yahoo.fr, karim.bouzoubaa@emi.ac.ma, prosso@dsic.upv.es

Abstract

The development of sophisticated applications in the field of the Arabic Natural Language Processing (ANLP) depends on the availability of resources. In the context of previous works related to the domain of the Arabic Question/Answering (Q/A) systems, a semantic Query Expansion approach using Arabic WordNet (AWN) has been evaluated. The obtained results, although AWN (one of the rare resources) has a low coverage of the Arabic language, showed that it helps to improve performances. The evaluation process integrates a Passage Retrieval (PR) system which helps to rank the returned passages according to their structure similarity with the question. In this paper, we investigate the usefulness of enriching AWN by means of the Yago ontology. Preliminary experiments show that this technique helps to extend and improve the processed questions.

1. Introduction

Arabic Natural Language Processing (ANLP) has known interesting attempts in the last years especially in morphology and less advanced Information Retrieval (IR) systems. However, the development of more sophisticated applications such as Question/Answering (Q/A), Search Engines (SEs) and Machine Translation (MT) has still a common problem: the lack of available electronic resources.

The Arabic WordNet (AWN) ontology (Elkateb et al., 2006) is one of these few resources. The AWN[1] is a lexical ontology composed of 23,000 Arabic words and 10 thousands of synsets (sets of words having a common meaning). The design of AWN presents many advantages for its use in the context of ANLP. Indeed, AWN has the same structure as the Princeton WordNet (PWN) (Fellbaum, 2000) and WordNets of other languages. The AWN ontology is also a semantic resource since it contains relations between its synsets and links to the concepts of the Suggested Upper Model Ontology (SUMO) (Niles & Pease, 2003). The advantages described above show that AWN can contribute in the development of sophisticated applications as well as the development of cross-language systems.

In a previous work on Arabic Q/A, (Abouenour et al., 2009b) proposed a Query Expansion (QE) approach which relies on the AWN content and its semantic relations, namely synonymy, hypernymy, hyponymy and definition. The proposed approach has improved the performances in terms of accuracy, the MRR[2] and the number of answered questions.

The reached performances have been considered encouraging for the following reasons:

- a number of 2,264 of well-known question sets in the field of Q/A and IR, namely the TREC[3] and CLEF[4] collections, were used;
- The difference of performances before and after using AWN is significant;
- Experiments have been conducted in an open domain (the web) which is a challenging context for Q/A systems.

Even though AWN has a low coverage of the Arabic language regarding other languages such as English, it helped to improve performances.

In order to enhance further performances, the idea is to develop and use a more enriched release of the AWN ontology. The enrichment of AWN could be done according to different lines: adding new synsets, enriching the existing synsets, enriching the hypoymy/hypernymy relations, verb categorization, Named Entity (NE) synsets, gloss, forms (for instance broken plurals), etc. The aim was focusing on the AWN lacks related to the used question collections. Therefore, the authors have performed an analysis of

[1] http://www.globalwordnet.org/AWN/

[2] Mean Reciprocal Rank (MRR) is defined as the average of the reciprocal ranks of the results for a sample of queries (the reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer).

[3] Text REtrieval Conference, http://trec.nist.gov/data/qa.html

[4] Cross Language Evaluation Forum, http://www.clef-campaign.org

the questions which contain keywords that can not be found in AWN (not extensible questions) and those for which the system could not reach the expected answer (not answered questions). For the two types of questions, they investigated either the keywords forming the questions and the type of the expected answer.

The analysis showed that for a high percentage of the considered questions, both the question keywords and answers are NEs. Hence, the enrichment of the NE content in the AWN ontology could help us to reach higher performances.

In this paper, we present an attempt to perform an automatic AWN enrichment for the NE synsets. Indeed, the use of a NER system (if such system is available and accurate in the context of the Arabic language) allows only identifying NE and information related to them whereas adding NE in AWN helps also to identify synsets which are semantically related to them (synonyms, subtypes, supertypes, etc.). Moreover, such enrichment could be also useful in the context of other ANLP and Cross-language tasks.

The current work is based on the Yago[5] ontology which contains 2 million entities (such as persons, organizations, cities, etc.). This ontology (Suchanek et al., 2007) contains 20 million facts about these entities. The main reasons behind using this resource are:

- its large coverage of NEs can help to improve performances in the context of Arabic Q/A systems;
- its connection to the PWN and the SUMO ontology (Gerard et al., 2008) can help us to transfer the large coverage of Yago to the AWN ontology.

The rest of the paper is structured as follows: Section 2 describes works using AWN; Section 3 presents the technique proposed for the AWN enrichment; Section 4 is devoted to the presentation of the preliminary experiments that we have conducted on the basis of the Yago content; in Section 5 we draw the main conclusions of this work and we discuss future work.

2. Arabic WordNet in previous works

There are many works that have integrated AWN as a lexical or a semantic resource. To our knowledge, most of these works belongs to the Arabic IR and Q/A fields. Indeed, in (El Amine, 2009), AWN has been used as a lexical resource for a QE process in the context of the IR task.

In the context of Q/A systems, authors in (Brini et al., 2009) have proposed an Arabic Q/A system called QASAL. They have reported that it will be necessary in future works to consider, the synonymy relations between AWN synsets at the question analysis stage of the proposed system. In (Benajiba et al., 2009), the authors have reported that the use of AWN would allow exploring the impact of semantic features for the Arabic Named Entity Recognition (NER) task which is generally included in the first question analysis step of a Q/A process (generally composed by three steps: question analysis, passages re-trieval and answer extraction).

In (Abouenour et al., 2008; Abouenour et al., 2009a), the authors have shown how it is possible to build an ontology for QE and semantic reasoning in the context of the Arabic Q/A task. In addition, the usefulness of AWN as a semantic resource for QE has been proved in the recent work of (Abouenour et al., 2009a) where the authors have considered not only the lexical side of AWN, but also its semantic and knowledge parts. Moreover, the QE process based on AWN has been used together with a structure-based technique for Passage Retrieval (PR). Indeed, the first step of our approach is retrieving a large number of passages which could contain the answer to the entered question. Generally, the answer is expected to appear in those passages nearer to the other keywords of the question or to the terms which are semantically related to those keywords. Therefore, new queries from the question were generated by replacing a keyword by its related terms in AWN regarding the four semantic relations mentioned previously.

In the second step of the described approach, the returned passages have to be ranked according to the structure similarity between the passages and the question. Thus, this step allows decreasing the number of passages to be processed at the answer extraction stage.

The conducted experiments showed an improvement of performances thanks to our two steps approach based on the AWN ontology for QE and the Java Information Retrieval System[6] (JIRS) (Gomez et al., 2007) for structure based PR. The analysis of the obtained results showed that:

- A high percentage (46.2%) of the TREC and CLEF questions are of NEs;
- The enrichment of the NE content in AWN will allow extending 69% of the non extensible questions;
- For a high percentage of the considered questions (50%), we can reach a similarity (between the question and passages) equal or higher than 0.9 and an average of 0.95 (max is 1) by using AWN together with JIRS.

[5] Yet an Other Great Ontology, available at http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

[6] http://jirs.dsic.upv.es

Thus, according to this analysis, the priority in terms of AWN enrichment is clear: in order to evaluate the QE and structure-based approach, we have to enlarge and refine the coverage, hierarchy and relations related to the NE synsets in AWN.

In the next section, we describe how resources belonging to other languages could be used for the enrichment of the NE content in AWN.

3. Enrichment of Arabic WordNet using Yago

According to the great number of words of the Modern Standard Arabic (MSA) language, the current release of AWN which has been manually built has still to be enlarged. The automatic enrichment is a promising way to reach a large coverage by AWN regarding the MSA. In this context, authors in (Al Khalifa and Rodriguez, 2009) have proposed a new approach for extending automatically the NE coverage of AWN. This approach relies on Wikipedia[7]. The evaluation done in that work shows that 93.3% of the NE synsets which was automatically recovered are correct. However, due to the small size of the Arabic wikipedia, only 3,854 Arabic NEs have been recovered.

Our approach proposes using a freely available ontology with a large coverage of NE instead of the Arabic Wikipedia. In addition to Yago, the field of open source ontologies provides interesting resources and attempts which belong either to the specific and open domain category: OpenCyc (Matuszek et al., 2006), Know-ItAll (Etzioni et al., 2004), HowNet[8], SNOMED[9], GeneOntology[10], etc.

For the purpose of the current work, we have been interested in using Yago for the following reasons (Suchanek et al., 2007):

- It covers a great amount of individuals (2 millions NEs),
- It has a near-human accuracy around 95%,
- It is built from WordNet and Wikipedia,
- It is connected with the SUMO ontology,
- It exists in many formats (XML, SQL, RDF, Notation 3, etc.) and is available with tools[11] which facilitate exporting and querying it.

The Yago ontology contains two types of information: entities and facts. The former are NE instances (from Wikipedia) and concepts (from WordNet), whereas the latter are facts which set a relation between these entities. To our knowledge Yago has been used as a semantic resource in the context of IR systems (Pound et al., 2009).

As we are interested in enriching the NE content of AWN, a translation stage has to be considered in our process. In (Al Khalifa and Rodriguez, 2009), authors used the Arabic counterpart of the English Wikipedia pages as a translation technique. In the current work, we consider instead the Google Translation API[12] (GTA) because its coverage for NEs written in Arabic is higher than the one of Arabic Wikipedia. In addition, translating a word using GTA is faster. Indeed, the result of a translation using Arabic Wikipedia needs to be disambiguated as many possible words are returned. This is not the case for the GTA.

The enrichment concerns both adding new individuals (NE) and adding their supertypes. These supertypes are very important and useful in our QE process combined to the structure-based PR system (JIRS). In order to show this usefulness, let us consider the example of the TREC question " متى ولد ليندون جونسون ؟" (When was Lindon Johnson born?). When we query a search engine using this question, the two following passages could be returned:

سنة 1908 و هو العام الذي **ولد** فيه ***ليندون جونسون*** ...	The year 1908 which is the year of birth of Lindon Johnson ...
ولد **الرئيس الأمريكي** ***ليندون جونسون*** يوم 27 أغسطس 1908 ...	The American president Lindon Johnson was born in 27 August 1908 ...

According to the two passages above, the JIRS system will consider the first passage as being the most relevant. Indeed, since the two passages contain the keywords of the question (ولد، ليندون جونسون), the similarity of the structure of each passage to the one of the question is the criterion to be used to compare them. The second passage contains a structure similar to the question with two additional terms (which are not among the question keywords) whereas in the first passage only one additional term appears (fyh - فيه). Therefore, the latter is considered more similar to the question than the former one. After enriching AWN by the NE ليندون جونسون and its supertypes such as رئيس أمريكي (r}ys >mryky : US President), we can consider, in the query processed by JIRS, the extended form of the question where the NE is preceded by its supertype الرئيس الأمريكي. In this case, the two terms الرئيس and الأمريكي are considered as being among the question keywords. Hence, the structure of the second passage would then be considered by JIRS as the most similar to the structure of the question. The second passage is the one containing the expected answer in a structure which

[7] www.wikipedia.org/

[8] www.keenage.com/html/e_index.html

[9] www.snomed.org

[10] www.geneontology.org

[11] http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

[12] http://code.google.com/p/google-api-translate-java/

structure which can be easy to process by the answer extraction module. In order to enrich the NE content in AWN, we have adopted an approach composed of seven steps. Figure 1 below illustrates these steps.

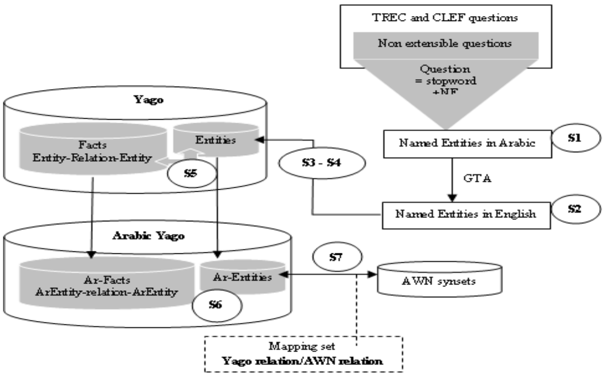

Figure 1: Steps of our approach for the enrichment of the NE content in AWN

As we can see, for the purpose of the current paper, we are interested in the enrichment of the NE part of AWN for the not extensible questions (547 TREC and CLEF questions). In order to do so, our approach relies on the following steps:

(S1) For each considered question, we extract the NE in Arabic;
(S2) Using the GTA, we translate the NE into English (GTA performs a disambiguation process);
(S3) We extract the Yago Entity related to the NE translated into English;
(S4) We extract the Yago Facts related to the Yago Entity extracted in the previous step;
(S5) In this step, we have a sub release of Yago related to the considered questions;
(S6) Using the GTA, we translate the content (entities and facts) of the sub release of Yago built in step five;
(S7) We perform a mapping between the NEs contained in the Arabic Yago (of step S4) and their related entries in AWN according to synonymy, hypernymy, hyponymy and SUMO relations.

After performing these steps, we have an enriched release of AWN which we consider in our new experiments. The obtained results in the enrichment and experimental processes are described in the next section.

4. Preliminary Experiments

As we have mentioned in the previous section, our focus is devoted to the NEs which appear in the not extensible questions. The number of these questions is 547. There are some NEs which appear in many questions. The number of distinct NEs is 472.

After performing steps 3 and 4, 374 distinct NEs (79%) have been identified within the Yago ontology. A number of 59,747 facts concern the identified Yago entities, with an average of 160 facts per entity. The average of the confidence related to these facts around 0.97 (the max is 1). The Yago ontology contains 96 relations. We have identified 43 relations in the facts corresponding to the NEs extracted from the considered questions. The TYPE relation is the first one to be considered in our approach for the enrichment of NEs in the AWN. For the purpose of the current work, we have considered only the facts containing a TYPE relation between a Yago entity and a WordNet concept. From the 374 NEs identified in Yago, 204 of them (around 55%) have a TYPE relation with a WordNet concept.

Relying on these relations on one hand and on the relation between the AWN synsets and the WordNet synsets on the other hand, we were able to connect 189 Yago entities (roughly 51% of the NEs of the considered questions) with the corresponding AWN synsets.

In order to connect the rest of NEs (185) with the AWN synsets (102 distinct synsets), we have set, in the context of the step S7 mentioned previously, different mappings between the relations used in the Yago facts and the corresponding AWN synsets. For instance, the second arguments of the relations "**citizenOf**", "**livesIn**", "**bornIn**", "**hasCapital**" or "**locatedIn**" are candidate hyponyms of the AWN synset "مدينة" (mdynp : city).

The enriched release of AWN that we have built using Yago helped us extending more questions and conducting preliminary experiments in the same way of (Abouenour et al., 2009a). Table 1 shows the obtained results.

Measures	before Yago	Using Yago
Accuracy	17,49%	23,53%
MRR	7,98	9,59
Number answered questions	23,15%	31,37%

Table 1: Results of preliminary experiments related the non extensible questions.

As we can see, performances in terms of accuracy, MRR and the number of answered questions have been improved after using our semantic QE which relies on the AWN release enriched with Yago.

5. Conclusion and Future Works

In this paper, we have proposed an approach to enrich AWN from the available content of the Yago ontology. The enrichment process was possible thanks to the connection existing between Yago entities and

WordNet on one hand and between WordNet and AWN on the other hand. In the preliminary experiments that we have conducted, we have considered the previous semantic QE approach which relies now on the new content of AWN. These experiments show an improvement in terms of accuracy, MRR and the number of answered questions.

In the current work, we have considered only the relations of Yago which allow a direct mapping between its entities and the AWN synsets. Therefore, considering the other relations and the whole content of Yago is among the intended future works.

Acknowledgement

This research work is the result of the collaboration in the framework of the bilateral Spain-Morocco AECID-PCI C/026728/09 research project. The third author thanks also the TIN2009-13391-C04-03 research project.

References

Al Khalifa M. and Rodríguez H. 2009. “Automatically Extending NE coverage of Arabic WordNet using Wikipedia”. In Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco, May, 2009.

Abouenour L., Bouzoubaa K., Rosso P., 2009. “Three-level approach for Passage Retrieval in Arabic Question /Answering Systems”. In Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco, May, 2009.

Abouenour L., Bouzoubaa K., Rosso P., 2009. “Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system” . In: Proc. Workshop on Computational Approaches to Semitic Languages, E-ACL-2009, Athens, Greece.

Abouenour L., Bouzoubaa K., Rosso P. 2008. Improving Q/A Using Arabic Wordnet. In: Proc. *The 2008 International Arab Conference on Information Technology (ACIT'2008),*Tunisia, December.

Benajiba Y., Mona D., Rosso P. Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. In: IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Processing Morphologically Rich Languages, Vol. 17, No. 5, July 2009.

Brini W., Ellouze M., Hadrich Belguith L. 2009. *QA SAL :* “Un système de question-réponse dédié pour les questions factuelles en langue Arabe*”. In: 9ème Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique, Tunisia.* (in French).

El Amine M. A. 2009. Vers une interface pour l’enrichissement des requêtes en arabe dans un système de recherche d’information. In Proceedings of the 2nd Conférence Internationale sur l'informatique et ses Applications (CIIA'09) Saida, Algeria, May 3-4, 2009.

Elkateb S., Black W., Vossen P., Farwell D., Rodríguez H., Pease A., Alkhalifa M. 2006. “Arabic WordNet and the Challenges of Arabic”. *In proceedings of Arabic NLP/MT Conference*, London, U.K.

Etzioni O., M. J. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. 2004. Web-scale information extraction in KnowItAll. In WWW, 2004.

Fellbaum C. 2000. “WordNet: An Electronic Lexical Database”. *MIT Press*, *cogsci.princeton.edu/˜wn,* September 7.

Gerard D. M., Suchanek F. M., Pease A. Integrating YAGO into the Suggested Upper Merged Ontology. 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008). Dayton, Ohio, USA (2008).

Gómez J. M., Rosso P., Sanchis E. 2007. Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system. In: Proc. Workshop on Cross Lingual Information Access, CLIA-2007, 20th Int. Joint Conf. on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12.

Matuszek C., J. Cabral, M. Witbrock, and J. De Oliveira. An introduction to the syntax and content of Cyc. In AAAI Spring Symposium, 2006.

Niles I., Pease A. 2003. “Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology.” *In Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada.

Pound J., Ihab F. I., and Weddell. G. 2009. QUICK: Queries Using Inferred Concepts fromKeywords Technical Report CS-2009-18. Waterlo, Canada.

Rodríguez H., Farwell D., Farreres J., Bertran M., Alkhalifa M., Antonia Martí M., Black W., Elkateb S., Kirk J., Pease A., Vossen P., and Fellbaum C. 2008. Arabic WordNet: Current State and Future Extensions in: Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008.

Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In Proc. of the 16th WWW, pp. 697-706 (2007).