ON THE EXTENSION OF ARABIC WORDNET NAMED ENTITIES AND ITS IMPACT ON QUESTION / ANSWERING

Lahsen Abouenour, Karim Bouzoubaa
Mohammadia School of Engineers, Med V Universit-Agdal, Agdal, Rabat, Morocco

Paolo Rosso
Natural Language Engineering Lab., ELiRF, Dpto. Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Valencia, Spain

Keywords: Arabic WordNet ontology, Question/Answering Systems, YAGO ontology, Named Entities, Semantic Query Expansion, Passage Retrieval.

Abstract: Applying a semantic approach has improved performances in the context of the Arabic Question/Answering (Q/A) systems. This approach is based on the current release of the Arabic WordNet (AWN) ontology. The analysis of the results obtained in previous works has shown that extending the Named Entity (NE) content of this ontology is a promising technique in order to further improve performances. In this paper, we investigate through experimental results the impact of the AWN enrichment using the YAGO ontology which presents a large coverage in terms of NE.

1 INTRODUCTION

As the construction of ontologies can be done using Natural Language Processing (NLP) tools (Charlet et al., 2009), there are also works related to NLP where ontologies are used as semantic resources. Indeed, researchers in the fields of NLP have devoted more interest in integrating ontologies in their works especially those belonging to particular domains such as Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT) and Question/Aswering (Q/A) (Buitelaar and Simiano, 2006).

The usefulness of ontologies in NLP should be evaluated with respect to a given task, resource and/or language. In IR and Q/A systems, ontologies are used mainly as resources for Query Expansion (QE) process (Voorhees, 1994), question analysis and answer extraction modules (Lopez and al., 2009). This is due to the advantages presented by ontologies and knowledge bases in terms of conceptualization and semantic information.

In the context of the Arabic language, the Arabic WordNet (AWN) ontology (Elkateb et al., 2006) remains among the few available resources which can be integrated in order to perform lexical and semantic processing. In recent works (Abouenour et al., 2009b), we have shown by experiments that integrating a QE process (Abouenour et al., 2008) based on the AWN ontology within an Arabic Q/A module improves the relevance of returned passages. Moreover, we have shown through a detailed example that performances can be enhanced even more by considering a similarity score calculated on the basis of semantic relations in the AWN ontology and Conceptual Graphs[1] (CG) representations. However, this promising approach could not be applied to a high number of the considered TREC[2] and CLEF[3] questions due to the low coverage of the AWN ontology. Hence, despite its high accuracy which is a feature of manually built resources, this ontology has to be enriched in order to benefit from its advantages in NLP works.

In (Abouenour et al., 2010a), we have analyzed the questions for which the proposed semantic

[1] A Conceptual Graph is a directed graph of nodes that correspond to concepts, connected by labelled and oriented arcs that represent conceptual relations (Sowa, 1984).

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

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

Abouenour L., Bouzoubaa K. and Rosso P..
ON THE EXTENSION OF ARABIC WORDNET NAMED ENTITIES AND ITS IMPACT ON QUESTION / ANSWERING.
DOI: 10.5220/0003102004240429
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 424-429
ISBN: 978-989-8425-29-4

approach could not be applied. This analysis showed that a high number of these questions are formed of Named Entities (NEs) or are expecting NEs (such as persons and organization names) as answers. Therefore, in that work, two tasks have been done on the basis of this analysis:

- extending the NE content in AWN which is related to the questions that could not be processed using the proposed semantic approach;
- conducting preliminary experiments in order to evaluate performances after this enrichment.

The first task has been performed by translating a small part of the YAGO[4] ontology content (Suchanek et al., 2007) to the Arabic language and by mapping this content to the AWN ontology synsets. At the end of this task, we were able to apply the semantic approach on the non processed questions (547 in number). The number of YAGO entities concerned in this step is negligible (374 out of 3 millions). Although AWN has been enriched with a very small number of NEs, preliminary experiments have shown that performances in terms of accuracy, MRR and the number of answered questions have been improved by 6.04%, 1,61 and 8,22% respectively.

In this paper, we conduct experiments in order to evaluate the impact of enriching the NE content of AWN using all available data of the YAGO ontology. Therefore, instead of considering only YAGO's entities and facts related to the non processed questions, we translate, firstly, all the content of this ontology and perform, secondly, a mapping between YAGO entities and AWN synsets.

The rest of the paper is structured as follows: Section 2 describes the features of the AWN ontology, its use in the context of Arabic NLP and previous works aiming its extension; Section 3 is devoted to the description of our technique using the YAGO ontology. The results of the experiments that we have conducted are presented and discussed in Section 4 and 5 respectively. Finally, in Section 6, we draw the main conclusions of the current work.

2 FEATURES AND AUTOMATIC ENRICHMENT OF AWN

Recent years have seen two significant facts: first we note an extension of information content particularly those produced by blogs and wikis. Moreover, with respect to linguistic resources that can be used in NLP, there is huge need of electronic dictionaries, lexical databases, ontologies, etc. This need is due to the fact that manually built resources with a high coverage are not freely available.

For languages such as English, there are many interesting open source ontologies which belong either to specific and open domain categories: OpenCyc (Matuszek et al., 2006), Know-ItAll (Etzioni et al., 2004), HowNet[5], SNOMED[6], GeneOntology[7], etc. To our knowledge, AWN ontology is the only freely available ontology in the context of the Arabic language. This lexical ontology is composed of 23,000 Arabic words and 10,000 thousands of synsets (sets of synonyms).

The interest devoted to AWN is due mainly to its compliancy and mapping with WordNets of other languages such as English (Fellbaum, 2000). The development of these WordNets is the subject of a periodical conference[8]. The AWN ontology is also connected to the Suggested Upper Model Ontology (SUMO) (Niles & Pease, 2003). AWN has been used in many Arabic NLP works such as Q/A systems (Brini et al., 2009), Arabic Named Entity Recognition (Benajiba et al., 2009), QE processes (El Amine, 2009), etc. These works showed that the techniques based on AWN are promising in the considered fields but the coverage of this ontology has to be improved in order to get higher performances.

A way to tackle this kind of problem is to use existing information content in order to improve the AWN coverage. Authors of (Al Khalifa and Rodriguez, 2009) have shown that using the Arabic Wikipedia[9] it is possible to enrich NEs in AWN. 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.

In (Abouenour et al., 2010a), we have proposed another technique aiming to enrich the NEs content in AWN. This technique relies on the YAGO ontology. Indeed, many advantages are presented by using YAGO such as its high coverage of NEs (3 millions), its accuracy of around 95%, its mapping

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

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

[6] www.snomed.org

[7] www.geneontology.org

[8] The last edition of the Global WordNet conference was held in February 2010 in Mumbai, India.

[9] www.wikipedia.org/

with WordNet, its connection with the SUMO ontology, its presentation in many standard formats such as RDF, etc (Suchanek et al., 2007). Let us recall briefly that the YAGO ontology is divided into two types of information: entities and facts. The former are NEs while the latter are relations between these NEs. This ontology has been recently used in many NLP works such as IR systems (Pound et al., 2009), automatic categories generation (Anjian et al., 2009), knowledge base building (Wang et al., 2010), etc.

In the next section, we present the technique used for automatically extending the NE content in YAGO.

3 EXTENTION TECHNIQUE USING YAGO

Identifying NEs within the Arabic Wikipedia content is one of the challenges reported in the work of (Al Khalifa and Rodriguez, 2009). The use of YAGO is, therefore, helpful in that NEs are already identified and checked in this ontology.

Figure 1 illustrates the different steps performed in order to enrich AWN entries based on YAGO entities.

As shown in Figure 1, the first step concerns the translation of all YAGO entities from English into the Arabic language. This translation is performed automatically using the Google Translation API (GTA). Generally, machine translation is more accurate in the case of NEs as they are formed of very few words.

The translated YAGO entities have been added in AWN according to two types of mappings as follows:

(i) The WordNet synsets corresponding to the YAGO entity are identified using the "TYPE" relation in the YAGO facts. After that, the AWN synsets corresponding to the identified WordNet synsets are connected with the given entity;

(ii) A mapping is performed between YAGO relations and AWN synsets. Table 1 shows some examples[10]. Based on this table, we extract the facts related to the mentioned YAGO relations. The second argument of the fact is candidate to be an instance of the corresponding AWN synset. In the case of the "isMarriedTo" both the first and the second arguments are considered. For the "inTimeZone" and "hasUnemployment" relations, we consider the first argument.

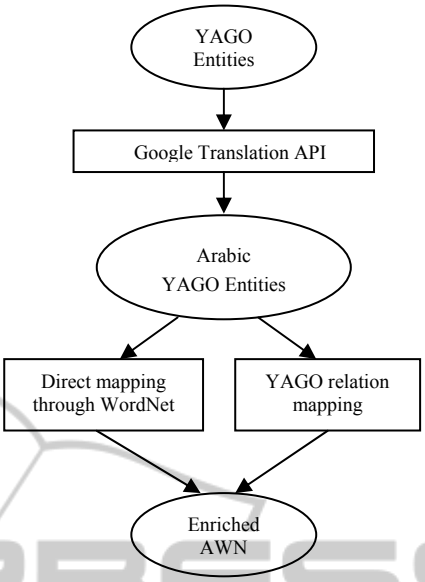

Figure 1: AWN ontology enrichment steps using YAGO.

In order to validate the final connections between YAGO entities (translated into Arabic) and the AWN synsets, we have eliminated some particular entities. Indeed, for each candidate entity, we query the Web using the exact expression formed from the synset and the candidate entity. If no document matches this query, we add the entity to a list for manual validation. For example, the YAGO entity "Wendy_Hall" appears in the following fact:

id: **401543569**
relation: **isLeaderOf**
argument 1: **Wendy_Hall**
argument 2: **Association_for_Computing_Machinery**

According to Table 1, the "Association for Computing Machinery" (جمعية الآلات الحاسبة) is candidate to be added to the AWN hierarchy under the "country" node as an instance. Using the Yahoo API, we search the content which matches the exact expression "بلد جمعية الآلات الحاسبة" (Association for Computing Machinery country).

Since there **is** no such content, we add this YAGO entity to the list of validation instead of adding it directly in the AWN ontology. This technique allows also eliminating some YAGO entities that have not been correctly translated. Table 2 shows the statistics after applying this technique.

[10] We use the Buckwalter transliteration schema to show romanized Arabic. This schema is widely used by the NLP community.

Table 1: Mapping between YAGO relation and AWN synsets.

YAGO relation	AWN synset	AWN synset id
actedIn	إبداع (creation : AibodaAE)	madiynap_n1AR
bornIn	مدينة (city : mdynp)	madiynap_n1AR
diedIn	مدينة (city : mdynp)	madiynap_n1AR
hasCapital	مدينة (city : mdynp)	madiynap_n1AR
hasCurrency	بلد (country : balad)	balad_n1AR
hasNumberOfPeople	بلد (country : balad)	balad_n1AR
hasPopulation	بلد (country : balad)	balad_n1AR
hasPopulationDensity	بلد (country : balad)	balad_n1AR
hasUnemployment	بلد (country : balad)	balad_n1AR
inTimeZone	منطقة-مقاطعة (region : mnTqp)	minoTaqap_n1AR
isCitizenOf	مدينة (city : mdynp)	madiynap_n1AR
isLeaderOf	بلد (country : balad)	balad_n1AR
isMarriedTo	زوج-زوجة (married : zwj)	zawoj_n1AR
livesIn	مدينة (city : mdynp)	madiynap_n1AR
locatedIn	مدينة (city : mdynp)	madiynap_n1AR
originatesFrom	منطقة (region : mnTqp)	minoTaqap_n1AR
politicianOf	بلد (country : balad)	balad_n1AR
worksAt	مُؤَسَّسَة (institution/establishment : mu&as~asap)	mu&as~asap_n1AR
wrote	كاتب (writer/author : kAtb)	kaAtib_n1AR

Table 2: YAGO and AWN evident mapping statistics.

YAGO relation	# entities	Eliminated entities
actedIn	28,836	35.09%
bornIn	36,189	20.59%
diedIn	13,618	12.92%
hasCapital	1,368	6.78%
hasCurrency	367	0.00%
hasNumberOfPeople	6,171	0.00%
hasPopulation	77,928	9.78%
hasPopulationDensity	44,628	0.00%
hasUnemployment	41	0.00%
inTimeZone	2	0.00%
isCitizenOf	4,865	0.00%
isLeaderOf	2,886	0.00%
isMarriedTo	8,416	0.00%
livesIn	14,710	11.11%
locatedIn	60,261	14.03%
originatesFrom	11,497	26.67%
politicianOf	6,198	0.00%
worksAt	1,401	3.45%
wrote	12,469	27.27%

As we can see, the three YAGO relations "wrote", "originatesFrom" and "bornIn" are the most concerned by the elimination of the entities to be added in AWN. In the case of the first relation, this is due to translation errors. For the second and third relations, the eliminated entities belong to cities or countries when we expect a region and to countries when we expect cities.

The number of entities that can be added to the AWN ontology using automatic mappings is 288,204. This is a very high number compared to the number of NEs in the current release of AWN (546). Note that we have not considered other YAGO entities due to several reasons (redundant entities, incorrect translation, etc.). As previously shown (Abouenour et al., 2010a), we expect to further improve performances with this large number of NEs. In the next section, we present the experimental results related to the evaluation that we have done in order to investigate this assumption.

4 EXPERIMENTAL RESULTS

In order to measure the impact of the automatic enrichment presented in the previous section, we have considered all the 2,264 TREC and CLEF questions translated into Arabic.

These experiments have been conducted in the same way as in (Abouenour et al., 2010a). Indeed, the semantic QE using AWN and the structure-based PR using the Java Information Retrieval System[11] (JIRS) (Gomez et al., 2007) are applied together on this set of questions. Accuracy, Mean Reciprocal Rank (MRR) and the number of answered questions have been used in order to measure the performance (Abouenour et al., 2010b). Table 3 shows the obtained results before and after using the enriched release of AWN.

Table 3: Results before and after AWN enrichment.

Measures	before YAGO	using YAGO
Accuracy	17.49%	25.22%
MRR	7.98	14.78
Number of answered questions	23.15%	35.05%

As we can see thanks to the enrichment of AWN using the YAGO ontology, performances have been significantly improved in terms of Accuracy, MRR and the number of answered questions by 7.73%, 6.80 and 11.90% respectively.

5 DISCUSSION

The enriched content of the AWN ontology allowed

[11] http://sourceforge.net/projects/jirs

us to apply the keyword-based and the structure-based approach on the TREC and CLEF questions which could not be processed previously. Moreover, performances have been enhanced even more according to the considered measures.

Improving the NE coverage of the AWN ontology had a positive impact on the effectiveness of the considered approach. These results confirm those obtained in preliminary experiments (Abouenour et al., 2010a) where only a restricted part of YAGO has been used. In that work, the accuracy has been improved from 17.49% to 23.53% (now it is 25.22%), the MRR from 7.98 to 9.59 (now it is 14.78) and the number of answered questions from 23.15% to 31.37% (now it is 35.05%).

These performances are due to the high coverage of YAGO in terms of names of Locations, Time and Persons existing in the considered questions (respectively 43%, 55% and 32.56%).

6 CONCLUSIONS

In this paper, we have shown through experiments that extending the coverage of NE in the AWN ontology using the YAGO ontology has a positive impact in the context of the Arabic Q/A systems. Indeed, applying a semantic QE approach based on the enriched content of AWN on a set of 2,264 TREC and CLEF questions allowed to reach better performances in terms of accuracy, MRR and the number of answered questions.

As future work, we plan to investigate the impact of enriching other main facets of AWN such as nouns and verbs using specialized ontologies and/or lexicons.

ACKNOWLEDGEMENTS

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 MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

REFERENCES

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 CITALA'09, Rabat, Morocco, May, 2009.

Abouenour L., Bouzoubaa K., Rosso P., 2010. Using the YAGO ontology as a resource for the enrichment of Named Entities in Arabic WordNet. Workshop LR & HLT for semitic languages, LREC'10. Malta. May 2010.

Abouenour L., Bouzoubaa K., Rosso P., 2010. An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. Special Issue in the International *Journal on Information and Communication Technologies/IEEE*. To appear in 2010.

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, March, 2009.

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.

Al Khalifa M. and Rodríguez R. Automatically Extending NE coverage of Arabic WordNet using Wikipedia. In Proc. of the *3rd International Conference on Arabic Language Processing* CITALA'09, Rabat, Morocco, May, 2009.

Anjian R, Du X., Wang P. 2009. Ontology-Based Categorization of Web Search Results Using YAGO. Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization. Vol. 01, pp 800-804. *IEEE Computer Society Washington*, DC, USA.

Benajiba Y., Diab M., Rosso P. 2009. 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.

Benoît S. and Darja F. 2008. Building a free French wordnet from multilingual resources. Workshop on Ontolex 2008, LREC'08. Marrakech, Maroc, June 2008.

Brini W., Ellouze M., Hadrich Belguith L. 2009. QASAL : 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*.

Buitelaar P. and Cimiano P. 2006. Ontology learning from text. Tutorial at EACL 2006. Trento, Italy; April, 2006.

Charlet J., Szulman S., Aussenac-Gilles N., Nazarenko A., Hernandez N., Nada N., Sardet E., Delahousse J. and Pierra G.. 2009. Apport des outils de TAL à la construction d'ontologies : propositions au sein de la plateforme DaFOE (poster). Dans *Journées*

Francophones d'Ingénierie des Connaissances (IC 2009), Hammamet, 25/05/2009-29/05/2009, Fabien Gandon (Eds.), INRIA, May 2009.

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, New York, NY, USA, May 2004.

Fellbaum C. 2000. WordNet: An Electronic Lexical Database. MIT Press, September, 2000.

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

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.

Lopez, V., Sabou M., Uren V. and Motta E. 2009. Cross-Ontology Question Answering on the Semantic Web – an initial evaluation, *Knowledge Capture Conference, 2009*, California.

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

Sowa J. F. 1984. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley Company.

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

Voorhees E. M. 1994. Query expansion using lexical-semantic relations, *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, p.61-69, July, 1994, Dublin, Ireland.

Wang Y., Zhu M., Qu L., Spaniol M. and Weikum G. 2010. Timely YAGO: harvesting, querying, and visualizing temporal knowledge from Wikipedia. EDBT 2010: 697-700. Lausanne, Switzerland. 2010.