Constructing a Lexicon of Arabic-English Named Entity using SMT and Semantic Linked Data

Emna Hkiri, Souheyl Mallat, Mounir Zrigui and Mourad Mars
Faculty of Sciences of Monastir, University of Monastir, Tunisia

Abstract: *Named Entity Recognition (NER) is the problem of locating and categorizing atomic entities in a given text. In this work, we used DBpedia Linked datasets and combined existing open source tools to generate from a parallel corpus a bilingual lexicon of Named Entities (NE). To annotate NE in the monolingual English corpus, we used linked data entities by mapping them to Gate Gazetteers. In order to translate entities identified by the gate tool from the English corpus, we used moses, a Statistical Machine Translation (SMT) system. The construction of the Arabic-English NE lexicon is based on the results of moses translation. Our method is fully automatic and aims to help Natural Language Processing (NLP) tasks such as, Machine Translation (MT) information retrieval, text mining and question answering. Our lexicon contains 48753 pairs of Arabic-English NE, it is freely available for use by other researchers*.

Keywords: *NER, named entity translation, parallel Arabic-English lexicon, DBpedia, linked data entities, parallel corpus, SMT.*

Received April 1, 2015; accepted October 7, 2015

1. Introduction

Named Entity Recognition (NER) consists in recognizing and categorizing all the Named Entities (NE) in a given text, corresponding to two intuitive classes; proper names (persons, locations and organizations), numeric expressions (time, date, money and percent). In our case we are tackling this issue in the context of Arabic-English Statistical Machine Translation (SMT) project. In this work, we use different DBpedia[1] resources and different techniques to identify NE.

One of the major tasks in Machine Translation (MT) is Named Entity (NE) translation; it consists of identifying the right translation among a number of translations with the same names. Based on the DBpedia semantic data, we developed a system that automatically extracts and translates NE.

Semitic language, in general, and Arabic language in particular presents a challenge in the NER. This is due that a wide range of NER systems are based on the initial capitalization of names to detect the proper names and on the upper case letters to detect the acronyms. In contrast, Arabic language does not provide such orthographic distinction and suffer from the lack of resources for NE [16]. In this work, we try to fill this gap by an Arabic-English lexicon of NE.

The rest of this paper is structured as follows: in Section 2 we present a state of the art of Arabic NE recognition and translation task. Section 3 we describe the architecture of our proposed system. Experimental results and related evaluations are detailed in Section 4. Section 5 concludes the paper with directions for future work.

2. State of the Art

Written and spoken Arabic is considered as difficult language to apprehend in the domain of Natural Language Processing (NLP), it presents specific morphological, syntactic, phonetic and phonological features [8]. The difficulty occurred in a complex implementation of a theoretical model or prototype into a real system usable in large scale applications [5, 19]

In the last decade, NE research around the world has taken giant leaps taking advantages of the creation of large annotated corpora and effective machine learning algorithms for various languages [17]. However, lexical resources and annotated corpora started appearing recently in Arabic [11]. Several works has been done in NE recognition and translation in Arabic. Here we present a brief survey.

2.1. Arabic NER

The majority of Arabic NER system focused mainly on traditional classes. Some systems introduced additional classes and categories related to a specific domain [12, 13]. Several works applied statistical learning algorithms to Arabic NER. Abdul-Hamid and Darwish [1] applied support vector machines algorithm. Nezda *et al.* [20] as well as Benajiba and Zitouni [7] use maximum entropy models. Mohit *et al.* [18] use Perceptron.

A range of works on Arabic NER includes labelled datasets, creation of gazetteers, rule-based and also statistical systems. RENAR (Rule-based Arabic NER

[1] http://dbpedia.org

system) is a recent rule-based approach [22]. It is based on gazetteers and grammar recognition rules to build the lexicon entities. The second system is proposed by Oudah and Shaalan [21], it is a hybrid approach and combines rule-based approach with various statistical classifiers for extracting out of a set of NE classes.

Wikipedia has been the test-bed for few studies and systems of NER. Mohit *et al.* [18] extended classes of NE and developed a NER system using Arabic Wikipedia.

2.2. Arabic NE Translation

One of the major tasks in MT is NE translation that is identifying the right translation for a given source language NE. There have been some successful attempts on the Arabic NE translation. Benajiba *et al.* [6] directly translated the NE using automatic word alignments. Hassan *et al.* [10] used similarity metrics to extract bilingual named entity from comparable and parallel corpora. Fehri *et al.* [9] translated Arabic NE to English using NooJ toolkit. She was based on a new representation model covering the sport domain. Abdul Rauf and Schwenk [2] improved NE translation leveraging comparable corpora and dictionaries, which contains unknown words. Ling *et al.* [15] used web links to attain the translation of the NE. The basic idea of our method is similar to that of Ling *et al.* [15] and Hassan *et al.* [10] in terms of used resources. In fact it combines between both of them. We used web links such as DBpedia linked Data for NE detection and recognition and the Parallel corpora for the translation of the spotted NE.

3. Proposed system of NE Recognition and Translation

In this section we present the architecture of the NE system. The used resources and tools are of two kinds. One was meant for the automatic annotation of the NEs such as DBpedia, SPARQL and Gate. The other one was meant for the translation through Moses decoder toolkit [14]. Our proposed system is composed of three phases the first one consists of the NE recognition. The second incorporates NE translation and the last is for bilingual lexicon generation.

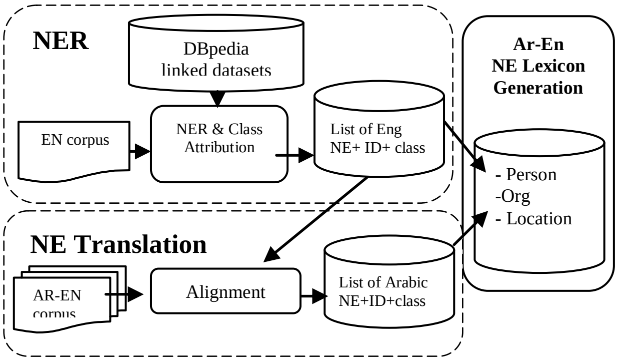

Figure 1. The architecture of our proposed system.

3.1. Named Entities Recognition

- *NE extraction*: Linked data is growing rapidly, now it includes more than 300 datasets, DBpedia has a very prominent and central place in this network. To extract the surface forms which are terms used to name the entities DBpedia provides three kind of sources: Data properties (DBpedia ontology properties), disambiguisation pages (dbont: wikiPageDisambiguates) and redirect pages (dbont:wikiPageRedirects). In our context we chose a short list of data properties that might contain surface form data: this list includes: rdfs:label foaf:name dbpprop:name. Therefore, to extract these entities, we interrogated the linked data set source using SPARQL Query. We saved 50 000 from each class of the spotted entities that are returned by the SPARQL query. These spotted entities (DBPentities) are in English language and then mapped to gazetteers for Gate tool.

Next, we add these new DBPentities to Gate's gazetteer as summarized in Table 1.

Table 1. NE extracted from DBpedia and mapped to Gate Gazetteers.

DBpedia entities	Gate Gazetteers	Overall
50 000 DBPPersons	10332 Persons	60332
50 000 DBPLocation	3434 Location	53434
50 000 DBPOrganization	3400 Organization	53400

- *Monolingual annotations*: Similar to the majority of the systems for Arabic and English NER we trained and evaluated our method leveraging corpora. In this step we used Gate and the mapped DBPentities to annotate the English corpus in order to extract our NE base containing persons, organization and location. We used 2000 and 2005 partitions of the United Nation (UN) corpus. After the elimination of replication and noise we obtained the set of NE as presented in Table 2.

Table 2. NE detected in the English corpus.

NE	2005 partition	2000 partition	Overall
DBPPersons	12645	21270	31642
DBPLocation	2882	2831	4471
DBPOrganization	10243	8653	18662

3.2. Named Entities Translation

- Corpus assessment: To obtain the corpus for our development, we downloaded the UN corpus for the years 2000/2005. Of this huge data, a portion for the years 2000 and 2005 were used for training and testing the system. The size of the experimental corpora is 1713134 sentences.

Before using our corpora for entities translation and extraction we applied some linguistic preprocessing.

1. *Tokenization*: is the first step in our method it aims to extract linguistic tokens from the graphic tokens.

Its output is the identification of token boundaries, punctuation marks and abbreviations; in addition to, the recognition of numbers which may be written in different alphabets.

2. *Truecasing*: is to convert initial word in each English sentence to its casing. This reduces data sparsity.
3. *Segmentation*: to analyze an Arabic text, we must run a segmentation phase because a paragraph may not contain any punctuation other than a point at the end. In addition, the Arabic sentence is long and complex; the recognition of the end is difficult because the punctuation is not systematic and sometimes particles can define another role than separating sentences.
4. *Data cleaning*: is the third step in our process, it is fundamental for obtaining a high quality of the MT output. In practice, it is hard to get a perfect or a near perfect data. By cleaning our training corpus, we removed:
   - The repetitive, misaligned or identical source-target segments.
   - Too short or too long segments or those that violate the ratio limit of Giza++.
   - Identified error patterns and web links (e-mail, ftp/ftps, http/https addresses).

The results of the linguistic preprocessing are presented in Table 3.

Table 3. Preprocessing results.

Language	English	Arabic
N° of tokens	47864964	40820597
N° of sentences	1713134	1713134
Average tokens/sentence	27,94	23,82

The results presented in the table show a difference in the number of tokens for Arabic and English corpus. This is an expected phenomenon because of the specificity of each language. The Arabic language is more complex than English and presents specific morphological, syntactic, phonetic and phonological features.

- *Building the language model*: We used the Arabic corpus to estimate the language model. We build a 5-gram Arabic language model using the SRILM Toolkit. This last is used to ensure fluent output and to find the smoothest translation among meaningful translations.
- *Training and Tuning*: To train our translation system we used parallel corpora (Arabic-English) which is aligned at the sentence level. The training was carried out in the 2000/2005 fields of the UN corpus: we used 1710000 sentences .For the tuning task we extracted aside the rest of the clean training corpus. We used about 3000 sentences. The tuning is the last task in our training process.

3.3. NE bilingual lexicon generation

The construction of the NE lexicon is based on the results of the SM translation. Once the English NE are recognized and translated in the English corpus a list of Arabic NE is generated.

The final output of this step consists of the aligned lists of NE (Arabic-English) indicating their class (person, organization and location) and given for each Arabic NE the same ID number of its corresponding English ID.

4. Evaluation

This section describes our evaluation results of the system described in the previous sections. We evaluated first the performance of our system to detect and recognize NE corresponding to person, location and organization classes. The second evaluation concerns the translation of the tagged NE to Arabic language based on parallel corpus. Finally, we compared our results with those of Attia *et al.* [4] and Mohit *et al.* [18].

4.1. Evaluation of the NER

In this step we evaluate our monolingual annotation of NE leveraging 2000/2005 fields. We kept a partition for evaluation which characteristics are presented in Table 4.

Table 4. Data collection.

	Evaluation corpus
Number of sentences	1150
Number of words	32131
Number of words/sent	27,94

The evaluation of our system is as follows:

- The texts of evaluation corpus are all manually annotated.
- After we have annotated automatically these texts using Gate and our system based on DBpedia Datasets.
- We have established a comparison between these annotations with corpus quality assurance. This Gate tool makes the comparison between two annotations on the same corpus.

The overall evaluation of entities extracted by our system is based on the use of F-measure. This measure combines the precision and the recall, it is defined as following (we have used the Fß=1.

$$F\_mesure = \frac{\left(\beta^2 +1\right) * precision * recall}{\beta^2 * precision + recall} \qquad (1)$$

Precision measurement is defined by the percentage of entities found by the system and which are correct. The recall is defined as the ratio between the numbers of found correct entities by the number of entities extracted from the reference articles. Table 5 shows results for different entities annotated by Gate using only the basic predefined gazetteers.

Table 5. Evaluation results of the NE recognition.

NE	Precision	Recall	F-measure
Person	83.04	79.34	81,14
Organization	78.6	62.11	69,38
Location	86.14	77.42	81,54
Overall	82,59	72,95	77,35

The strength of our Annotation system is in recognizing Location entities, which is attributed to the high named entity coverage of our NE base containing the DBpedia Datasets.

It is important to observe that our system shows good recall for Person Names which were more abundant in the parallel corpus. Besides this, the corpus had a heterogeneous mix of Person proper names not only from Arabic countries but also from the continents of Africa, Asia and America. A good recall percentage for Person entity names is encouraging as the NE related to south Asia and America have no phonetic similarity with the native person names entities available in Arabic countries. This gives good credence to our annotation process presented above

A detailed look to the results reveals that our NE translation system performs poorly for organizations entities annotation that appear in the corpus, our system cannot efficiently handle the acronyms short names or the abbreviations of these NEs.

These results are satisfying according to limitations due to the quality of the UN nation's corpus. As we said before we had several errors in the texts such as the existence of many paragraphs written in different languages such as Spanish, Italian, Dutch Russian etc., these errors change the precision and recall scores. In the next work, we plan to improve the quality of the texts (preprocessing and cleaning steps) also to increase the size of our corpus and the number of the DBpedia datasets to obtain a higher performance of the system

4.2. Evaluation of the NE Translation

This section presents our evaluation on the NE translation task. We compare the translation results obtained from our named entity translation system and human translations. The input of this task, as presented in the Figure1, is the set of Arabic and English NE pairs. The pair is the previously tagged Ne in the English corpus and candidate to NE selected, translated and tagged in the Arabic corpus. At the end, our system generates a list of suggested Arabic words {NE$_{ar}$, C, ID} for each English NE {NE$_{en}$, C, ID} with their ID -number

In order to evaluate the Arabic NE translation we pass by the following steps:

- We took arbitrarily pairs of NE in the lists {NE$_{ar}$,C, ID} and {NE$_{en}$, C, ID}
- Each NE was extracted, and paired with its translations using their ID -number in order to create the test set.
- The linguist manually translated all the NE in that test set. In total we obtained 1000 NE (person, location organization).
- Calculating both, the blue score and Out Of Vocabulary words (OOV). Table 6 presents the results that we obtained for the NE translation process.

Table 6. Evaluation results of the NE translation.

NE	Blue Score %	OOV%
Person	54.36	0.22
Organization	44.75	0.3
Location	57.33	0.21
Overall	52.14	0.24

Compared to the annotation task, the translation results are of poorer quality. This is due to several development and alignment issues. In fact, in time of development we observed that the training corpus contains several types of errors in sentences alignment and preprocessing. These errors in the training corpus affect to the SMT models and NE translation. In order to gain time, we have not made corrections of these errors. We reported them to another time of development. All the results and experiments presented in this phase are obtained using the training corpus without any corrections or modifications.

In order to obtain a complete and coherent bilingual lexicon of NE pairs, the Arabic translations were reviewed carefully to correct wrong translations made by our system. We used web pages to find the correct spelling especially for the abbreviations and acronyms in the organizations names. In cases where the Arabic translations could not be verified, the translations made by our linguist are considered as the correct translations. At the end our bilingual lexicon contains 48753 Pairs of Arabic-English NE as presented in Table 7.

Table 7. Parallel lexicon of Ar-EN names entities.

NE	Arabic-English Pairs
Person	27480
Organization	17237
Location	4036
Overall	48753

4.3. Comparison with Arabic Named Entity lexicon

We choose to compare our lexicon with that of Attia and Behrang because of our knowledge there is no free Arabic-English NE lexicon to date. Attia's lexicon 45000 NE; unfortunately the lexicon is not available online anymore, so we could not carry out a detailed comparison between the two lexicons. In his work, Attia *et al.* [4] built Arabic Named Entity lexicon based on web resources and linguistic toolkits. Behrang lexicon contains 75000 NE; but it is monolingual, it is free available same as our lexicon, also it takes advantage of linguistic corpus, tools and external web resources. Table 8 presents a comparison between our bilingual lexicon and those of Attia and Behrang.

Table 8. Comparison with Arabic NE lexicons.

	Our lexicon	Lexicon of Attia	Lexicon of Behrang
Bilingual	YES	NO	NO
Number of Arabic NE	65203	45000	75000
Number of English NE	65203	NO	NO
Freely available	Yes	NO	YES
NE classification	Yes	Yes	Yes
Linguistic tools	Moses Gate Giza++	MADA TOKAN MINELex	MADA ANer
External web resources	DBpedia Linked datsets	AWN WiKi GEONAMES	WIKI
Corpus	UN corpus	No	ACE &ANER corpus

As presented in the table above, the strength of our lexicon are the number and the quality of NE pairs; our lexicon contains 48753 bilingual, well structured and easy to exploit and freely available online[2]. It is a great resource which improves the quality of any Arabic English MT system [3]

5. Conclusions and Future Works

When dealing with NE Translation there are three main aspects that must be taken into account: the approaches to be used for recognizing and translating these NE and the characteristics of the language pair involved. To overcome these issues; we presented our NER and translation approach using DBpedia Linked datasets and parallel corpus. For annotating NE in monolingual English corpus we used Gate tool. Our approach is based on linked data entities by mapping them to Gate Gazetteers, and then constructing a type-oriented NE base covering person, Location and organization classes. The second task consists of translating these entities and then finally, generating our NE lexicon that encloses the list of Arabic entities that match to the English lists.

We made the bilingual lexicon available to the research community throughout the web. We believe NE lexicon is a very useful resource for NLP developers. We hope that free access to this lexicon will be beneficial for filling the gap of publicly available Arabic- English resources. We hope the release of this lexicon will further boost the research in the NLP field. Future works include development of a multilingual lexicon of NEs including other languages. We intend to increase the size of our DBpedia NE base in order to attain better performance for the annotation task. Also, to expand the size of our training data to reach better results in the translation step.

References

[1] Abdul-Hamid A. and Darwish K., "Simplified Feature Set for Arabic Named Entity Recognition," *in Proceeding of Named entities workshop. Association for Computational Linguistics*, Uppsala, pp. 110-115, 2010

[2] Abdul-Rauf S. and Schwenk H., "On the Use of Comparable Corpora to Improve SMT Performance," *in Proceeding of the 12th Conference of the European Chapter of the ACL*, Athens, pp. 16-23, 2009.

[3] Agrawal N. and Singla A., "Using Named Entity Recognition to Improve Machine Translation," Technical Report, 2012.

[4] Attia M., Toral A., Tounsi L., Monachini M., and Genabith J., "An Automatically Built Named Entity Lexicon for Arabic," *in Proceeding of International Conference on Language Resources and Evaluation*, Valletta, pp. 3614-3621, 2010.

[5] Ben Mohamed A., Mallat S., Nahdi M., and Zrigui M., "Exploring the Potential of Schemes in Building NLP Tools for Arabic Language," *The International Arab Journal of Information Technology*, vol. 12, no. 6, pp. 566-573, 2015.

[6] Benajiba Y., Rosso P., and Benedí Ruiz J., "ANERsys: an Arabic Named Entity Recognition System Based on Maximum Entropy," *in Proceeding of International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico, pp. 143-153, 2007.

[7] Benajiba Y. and Zitouni I., "Enhancing Mention Detection using Projection via Aligned Corpora," *in Proceeding of Conference on Empirical Methods in Natural Language Processing*, Massachusetts, pp. 993-1001, 2010.

[8] El-Jihad A., Yousfi A., and Si-Lhoussain A., "Morpho-Syntactic Tagging System Based on the Patterns Words for Arabic Texts," *The International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 350-354, 2011.

[9] Fehri H., Haddar K., and Ben Hamadou A., "Recognition and Translation of Arabic Named Entities with NooJ using a New Representation Model," *in Proceeding of 9th International Workshop on Finite State Methods and Natural Language Processing*, Blois, pp. 134-142, 2011.

[10] Hassan A., Fahmy H., and Hassan H., "Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora," *in Proceeding of Conference on Recent Advances in Natural Language Processing*, Borovets, pp. 1-6, 2007.

[2] The NE Lexicon and supplementary documents with details can be found at: https://github.com/Hkiri-Emna/Named_Entities_Lexicon_Project

[11] Hkiri E., Mallat S., and Zrigui M., "Automatic Translation of Arabic Texts Based on Ontology," *in Proceeding of International Conference on Web and Information Technologies*, Hammamet, pp. 494-501, 2013.

[12] Hkiri E., Mallat S., and Zrigui M., "Events Extraction From Arabic Text," *The International Journal of Information Retrieval Research*, vol. 6, no. 1, pp. 36-51, 2016.

[13] Hkiri E., Mallat S., Maraoui M., and Zrigui M., "Automating Event Recognition for SMT Systems," *in Proceeding of International Conference on Computational Collective Intelligence*, Madrid, pp. 494-502, 2015.

[14] Koehn P., Federico M., Cowan B., Zens R., Dyer C., Bojar O., Constantin A., and Herbst E., "Moses: Open Source Toolkit for Statistical Machine Translation," *in Proceeding of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, pp.177-180, 2007.

[15] Ling W., Calado P., Martins B., Trancoso I., Black A., and Coheur L., "Named Entity Translation using Anchor Texts," *in Proceeding of the International Workshop on Spoken Language* Translation, San Francisco, pp. 206-213, 2011.

[16] Mallat S., Ben Mohamed A., Hkiri E. , Zouaghi A., and Zrigui M., "Semantic and Contextual Knowledge Representation for Lexical Disambiguation: Case of Arabic-French Query Translation," *Journal of Computing and Information Technology*, vol. 22, no. 3, pp. 191-215, 2014.

[17] Mallat S., Hkiri E., Maraoui M., and Zrigui M., "Lexical Network Enrichment Using Association Rules Model," *in Proceeding of 16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, pp. 59-72, 2015.

[18] Mohit B., Schneider N., Bhowmick R., Oflazer K., and Smith N., "Recall-Oriented Learning of Named Entities in Arabic Wikipedia," *in Proceeding of 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, pp. 162-173, 2012.

[19] Nahar K., Al-Muhtaseb H., Al-Khatib W., Elshafei M., and Alghamdi M., "Arabic Phonemes Transcription using Data Driven Approach," *The International Arab Journal of Information Technology*, vol. 12, no. 3, pp. 237-245, 2015.

[20] Nezda L., Hickl A., Lehmann J., and Fayyaz S., "What in The World is a Shahab? Wide Coverage Named Entity Recognition for Arabic," *in Proceeding of International Conference on Language Resources and Evaluation*, Genoa, pp. 41-46, 2006.

[21] Oudah M. and Shaalan K., "A Pipeline Arabic Named Entity Recognition using a Hybrid Approach," *in Proceeding of International Conference on Computational Linguistics*, Mumbai, pp. 2159-2176, 2012.

[22] Zaghouani W., "RENAR: A rule-based Arabic named entity recognition system," *ACM Transactions Asian Language Information Processing*, vol. 11, no. 1, pp. 1-13, 2012.

Emna Hkiri is a PhD student in Computer Sciences at the Faculty of Economics and Management of Sfax. She is a member of LaTiCe Laboratory. His main research interests are in natural language processing (Arabic language); text translation, ontologies, NER and machine learning.

Souheyl mallat is a PhD student in the Faculty of Economic Sciences and Management of Sfax, Tunisia. He is member of LaTICE Laboratory, Monastir unity (Tunisia). His areas of interest include natural language processing, data mining and information retrieval.

Mounir Zrigui I received my PhD from the Paul Sabatier University, Toulouse, France in 1987and my HDR from the Stendhal University, Grenoble, France in 2008. Since 1986, I am a Computer Sciences assistant Professor in Brest University, France, and after in Faculty of Science of Monastir, Tunisia. I have started my research, focused on all aspects of automatic processing of natural language (written and oral), in RIADI laboratory and after in LaTICE Laboratory. I have run many research projects and published many research papers in reputed international journals/conferences.

Mourad Mars received his PhD from Grenoble Alpes University, France in 2011. He is member of LaTICE Laboratory, Monastir unity (Tunisia). His areas of interest include natural language processing, machine learning and features extracting.