Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Sellami, Rahma; Deffaf, Fatima; Sadat, Fatiha; Hadrich Belguith, Lamia
Improved Statistical Machine Translation by Cross-Linguistic Projection of Named Entities Recognition and Translation
Computación y Sistemas, vol. 19, núm. 4, 2015, pp. 701-711
Instituto Politécnico Nacional
Distrito Federal, México

ISSN 2007-9737

Improved Statistical Machine Translation by Cross-Linguistic Projection of Named Entities Recognition and Translation

Rahma Sellami[1], Fatima Deffaf[2], Fatiha Sadat[2], Lamia Hadrich Belguith[1]

[1] Sfax University, ANLP Research Group, MIRACL Laboratory, Sfax, Tunisia

[2] UQAM, Montreal, Canada

rahma.sellami@fsegs.rnu.tn, deffaf.fatima@courrier.uqam.ca, sadat.fatiha@uqam.ca, l.belguith@fsegs.rnu.tn

Abstract. One of the existing difficulties in natural language processing applications is the lack of appropriate tools for the recognition, translation, and/or transliteration of named entities (NEs), specifically for less-resourced languages. In this paper, we propose a new method to automatically label multilingual parallel data for Arabic-French pair of languages with named entity tags and build lexicons of those named entities with their transliteration and/or translation in the target language. For this purpose, we bring in a third well-resourced language, English, that might serve as pivot, in order to build an Arabic-French NE Translation lexicon. Evaluations on the Arabic-French pair of languages using English as pivot in the transitive model showed the effectiveness of the proposed method for mining Arabic-French named entities and their translations. Moreover, the integration of this component in statistical machine translation outperformed the baseline system.

Keywords. Named entity, pivot language, machine translation.

1 Introduction

Named Entities (NEs) are expressions commonly used and appearing frequently in all kinds of texts. NEs are very efficient elements in many Natural Language Processing (NLP) applications such as Cross-Language Information Retrieval (CLIR), Statistical Machine Translation (SMT) [22], [20], mention detection [24], news aggregation [16], and plagiarism detection [9]. Regularly updated documents such as news articles and Web pages usually contain a large number of proper names which are much more variable than common words and change continuously. This phenomenon is very problematic for the task of NE translation and affects directly the performance and quality of a phrase-based MT system. Hence, machine translation systems usually fail to capture those proper names. Moreover, a study of unknown words performed by Habash et al. [10] on Arabic into English translation showed that 40% of unknown words are proper names. Translating those proper names requires a specific treatment, especially in the case of multiwords. An example is "داني أبو لوح" (in Buckwalter transliteration: dAny Obw lwH[1]), that is an Arabic multiword representing a proper name. This NE is translated into English by many MT systems as "Danny Abu board"[2] instead of "Danny Abu Lwh", which is the correct translation. Zaghouani [23] listed a selection of freely available Arabic NE corpora. Only the JRC corpora contain Arabic-French NE translations. Thus, translating NEs is a challenging problem. Part of the reason is that NEs are either phonetically transliterated or semantically translated or both [20].

Due to the lack of appropriate resources for the task of NE recognition and translation for Arabic-French pair of languages, we bring in a third well-

[1] All Arabic transliterations are provided using the Buckwalter transliteration scheme [3].

[2] Translated with Google Translator on 1/20/2015.

resourced language, namely, English, which might serve as pivot, in order to build an Arabic-French NE translation lexicon.

First, the recognition of English NEs is completed using a parallel corpus on Arabic-English pair of languages. Second, we introduce a first translation component for those detected NEs in English into Arabic using aligned parallel corpora. Furthermore, a second translation component of those detected NEs in English into French is introduced using the phonetic similarities. Last, we merge those two translation components to build an Arabic-French NE translation lexicon.

A series of experiments have been conducted to estimate the performance of the proposed translation approach within SMT. Preliminary experimental results showed an improvement in terms of Bleu score, a decrease of the OOVs rate, and a better quality of translation after the integration of the proposed translation components into SMT.

This paper is organized as follows. In the next section, we review the related work on the detection of NE translations from parallel corpora. In Section 3, we present the proposed transitive model for the recognition and translation of Arabic-French NEs using English as a pivot language. We discuss our experimental setting and evaluations in Section 4. Section 5 concludes this paper and introduces some perspectives.

2 Related Work

As mentioned by Moore [17], two major strategies, symmetry and asymmetry, are used to automatically extract bilingual NEs from parallel corpora.

On the one hand, the symmetric strategy tries to find NEs in both languages and then to establish the associations between NE pairs. Chen et al. [4] analyzed the formulation and transformation rules for English-Chinese NEs. They used a frequency-based method to construct rules that identify NE keywords from phrase-aligned corpora. Huang et al. [11] proposed a method that acquires English-Chinese NE pairs from a parallel corpus, based on a linear combination of costs related to NEs transliteration, translation, and tagging. Kumano et al. [13] proposed a method for acquiring bilingual NE translations from non-literal content-aligned parallel corpora. First, they recognized the NEs in each of a bilingual document pair. Then they find NE groups whose members share the same referent. Finally, they completed a mapping between bilingual NE groups. Sellami et al. [20] introduced a framework to extract all NE translation types from a noisy parallel corpus. First, they recognized the NEs in each of a bilingual document pair. Second, they aligned the noisy parallel corpus on the sentence level. Then they performed a mapping between bilingual NE groups based on sentence alignment and NE type information. Finally, they filtered bad translations using statistic and linguistic information.

On the other hand, the asymmetric strategy assumes that NEs in the source part of the parallel corpus are given and that the main problem is related to the identification of their translation equivalents in the target part of the parallel corpus. Al-Onaizan and Knight [1] proposed a transliteration algorithm based on sound and spelling mapping using a finite state machine. Moore [17] proposed three progressively refined phrase translation models to learn the translations of NE phrases from parallel software manual corpus. These statistical models depend heavily on the linguistic information such as the same NE phrase occurring in the source and target parts and cues from capitalization. The obvious restriction on the applicability of the proposed models is that it requires the target language translations of source language phrases to be contiguous. Feng et al. [6] proposed a maximum entropy model that integrates the translation, transliteration, co-occurrence, and distortion scores in order to extract English-Chinese NE equivalents from a parallel corpus. Samy et al. [19] proposed a simple mapping scheme to transliterate Arabic NEs into Spanish based on an Arabic-Spanish parallel corpus and a Spanish NE tagger. First, they tag Spanish NEs. Second, they transliterate all the words in an Arabic sentence using a mapping scheme. Then, for each aligned pair of sentences, they consider words matching with Spanish NEs as Arabic transliteration. Lee et al. [14] introduced an approach that aligns bilingual NEs in parallel corpora by incorporating statistical models with multiple knowledge

ISSN 2007-9737

sources. They modeled the process of translating an English NE phrase into a Chinese equivalent using lexical translation/transliteration probabilities for word translation and alignment probabilities for word reordering. Azab et al. [2] proposed a model that extracts and labels parallel NEs from a large English-Arabic parallel corpus. They started by tagging the English sentences with NE classes. Then, they used word alignment to project and collect the associated Arabic NEs. To reduce the noisy nature of word alignments, they designed a procedure to clean up the noisy Arabic NE spans by POS verification. Finally, Darwish et al. [5] proposed a generative model for transliteration mining from Wikipedia inter-wiki-link data.

Obviously, the symmetric approach is more difficult to apply, due to the lack or the performance of NE recognition tools for many languages. Moreover, the errors and inconsistency induced by NE identification, subsequently, infect the NE alignment [20].

After this review of previous works, it is clear that most works used the English language as source or target language in the process of NE transliteration; this is due to the lack of resources, such as parallel corpora, NE recognizer, transliteration tool, etc., for languages other than English.

In this paper, we introduce a new asymetric strategy based on a transitive model for the recognition of Arabic NEs and their translation into French using English as a pivot language. Our proposed model does not require tools for each of the source (Arabic) or target language (French), such as Arabic-French parallel corpora, or Arabic NE recognizer, or French NE recognizer.

3 A Transitive Model for Arabic-French NE Recognition and Translation

In this section, we describe our method for the recognition of Arabic NEs and their translation into French using English as a pivot language. Figure 1 shows the process of the proposed method. First, English NEs in the English corpus are recognized. Second, English-Arabic and English-French NE lexicons are constructed using a cross-linguistic projection method based on parallel corpora of both pairs of languages. Later, a merging process of both NE lexicons based on English as a pivot language is performed in order to produce an Arabic-French NE lexicon.

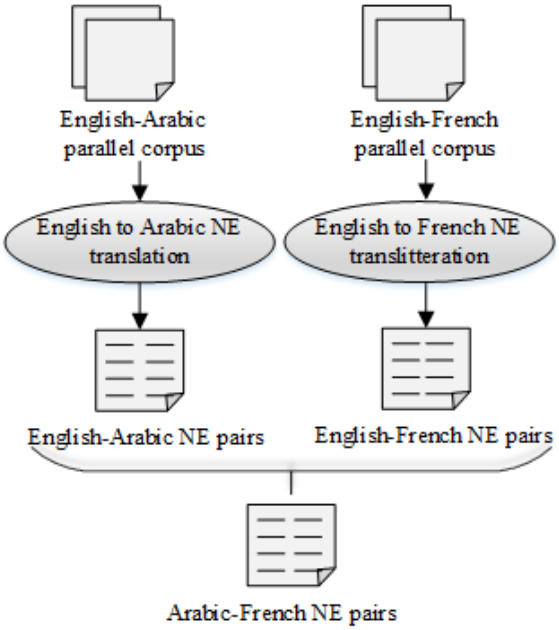

Fig. 1. The process of Arabic to French NE translation

3.1 English-Arabic NE Translation

This section explains the process of English into Arabic NE translation. The proposed method is very close to the proposal of Samy et al.[19], except we first transliterate English NEs and then we try to find correspondence in the Arabic parallel sentence whereas Samy et al. transliterate all the words in the Arabic corpus and then they try to find correspondence with Spanish NEs. Moreover, an Arabic normalization task is added in order to reduce the number of transliteration candidates. The translation process is simplistic, straightforward, and has shown its efficiency on the Spanish-Arabic pair of languages [19].

Our method is based on the proposed assumption of Samy et al. [19] as follows: "Given a pair of parallel sentences and given that in one sentence, one or more NEs were detected, then the corresponding aligned sentence should contain

the same transliterated NE". This assumption is a simplistic one, as it does not take into consideration common phenomena in translation such as omission or addition. Starting from this assumption, we describe our method for translating English NEs into Arabic.

The input consists of the aligned English-Arabic parallel corpus with a tagged English NE. The corpus is processed so that each pair of aligned sentences is handled one at a time. To avoid encoding scheme problems or unrecognized characters, we implemented numerical codification using the Unicode value for each Arabic character.

If an NE exists in the English sentence, the following steps are performed.

3.1.1 Arabic Sentence Pre-processing

Pre-processing Arabic sentences involves the process of tokenization, stop-words removal, and normalization. First, the process of tokenization involves simple punctuation splitting. However, NEs, like other nouns in Arabic, may be proceeded by clitics, such as the conjunction "و / w", prepositions "ب / b", "ل / l", or both "وب / wb", "ول / wl". To handle such a feature, we had to expand the possibilities of matching by indicating that the string might be preceded by one or more clitics. Second, a stop word filter excludes the stop words from the potential candidate translation. Finally, a normalization process that consists in grouping together similar letters in Arabic in such a way that letters in one group correspond to similar pronunciations and then representing each group with one letter. For example, alef maqsura "ى / Y" is converted into Yaa "ي / y" and "ش / $", "ص / S", "ز / z" are converted into "س / s". The normalization process includes also removing kashida and short vowels.

Based on the normalization process, the transliterated Unicode code is normalized in order to reduce the number of transliteration possibilities. Table 1 shows some examples of the normalization process.

The following steps describe the transliteration process of English NEs into the normalized Arabic Unicode code.

3.1.2 Single Word NE Transliteration

For each English NE and according to a mapping scheme, the system provides a combination of all possibilities of transliteration into the Arabic Unicode code. Table 2 shows some examples of mapping English letters and Unicode codes.

The output consists of the English NEs together with their transliteration hypothesis. The process of transliteration of an English NE is completed as follows:

1. For each English sentence in the English-Arabic parallel corpus, detect all English NEs using Stanford NER [3] [7]. The Stanford NE Recognizer is a CRF Classifier. The classifier was trained on the CoNLL 2003[4] English training data.
2. For each English NE, generate all Arabic NE candidates using the Unicode transliteration scheme.
3. If the Arabic NE candidate matches any term in the aligned Arabic sentence of the target side of the English-Arabic parallel corpus, then consider the Arabic NE candidate as the best transliteration of the English NE. Otherwise, consider the English NE as unknown word.

An example is the English proper name "Nidal" to which the transliteration module generates successively the Arabic transliteration candidates as shown in Table 3. When comparing each Arabic NE transliteration candidate with each term in the Arabic sentence of the Arabic side of the parallel corpus, we consider the candidate as a transliteration hypothesis according to the following two rules:

1. The transliteration corresponds exactly or is similar to a term in the Arabic sentence with less than two different characters.
2. The transliteration corresponds or is similar to the concatenation of multiple terms in the Arabic sentence.

[3] http://nlp.stanford.edu/software/CRF-NER.shtml
[4] http://www.cnts.ua.ac.be/conll2003/ner/

Table 1. Examples of normalization

Arabic NE	Letters to normalize	Normalized NE
السيد أنصارى /Alsyd >nSArY	أ، ى، ص	السيد انساري /Alsyd AnsAry
نضال / nDAl	ض	ندال / ndAl
سوشاريبا / sw$ArybA	ش	سوساريبا / swsArybA

Table 2. Example of mapping scheme

English letter	The Unicode code
N	\u0646 (ن)
I	Null, \u064A (ي),\u0627 (ا), \u0639 (ع), \u0627 \u064A (اي)
D	\u062F (د), \u062A (ت), Null
A	Null, \u0627 (ا), \u0624 (ي), \u062A (ت), \u0639 (ع), \u0627\u064A (??), \u0639\u0627 (عا)
L	\u0644 (ل), \u064A (ي), Null

3.1.3 Multiword NE Transliteration

A multiword NE consists of more than one word, for example, “Adam Smith”, “South Africa”, “Mohammed Rateb Nabulsi”. In this case, we perform the following steps:

1. Counting the number of words (N) composing the NE.
2. Transliterating the first word following the transliteration module of a single word NE.
3. If the first word corresponds to a term in the aligned Arabic sentence, then add the following N-1 Arabic terms to the term found in step 2.

Table 4 shows an example of a parallel sentence containing multiword NEs. The number of tokens in the English NE, N is equal to 2. The first word of the English NE is transliterated into the word “محمد/mHmd” witch exists in the Arabic sentence. Then the N-1 tokens following the word “محمد” are added. We obtain the transliterated multiword NE “محمد حسان/mHmd HsAn”.

3.1.4 Translation of Unknown NEs

Unknown NEs are English NEs whose Arabic equivalents cannot be fixed with the previous steps; such as the country name “Greece” which should be translated into “اليونان/AlywnAn”. These NEs are translated rather than transliterated. Such names failed to be recognized through the previous stages. In this case, a form of translation, either using a lexicon or machine translation, is required. Google Translator is used in order to translate those particular and few NEs among the long list of recognized English NEs; then a search for the Arabic NE is performed in the aligned Arabic sentence of the target side of the English-Arabic parallel corpus.

3.2 English to French NE Transliteration

In this section, we introduce our method for building an English-French NE transliteration lexicon based on the phonetic similarity. In Section 3.2.1, we introduce our method for English to French named entity transliteration. In Section 3.2.2, we describe the transliteration similarity measure used in this proposal.

Table 3. The Arabic transliteration candidates generated for the English NE "Nidal"

Iteration	Code of the candidate transliteration	Arabic string	Status
1	\u0646 NULL \u062F NULL \u0644	ندل	No
2	\u0646 NULL \u062F NULL \u064A	ندي	No
3	\u0646 NULL \u062F NULL NULL	ند	No
4	\u0646 NULL \u062F \u0627 \u0644	ندال	Yes

Table 4. An example of parallel sentences containing multiword NEs

English sentence	...Mohamed Hassen was the guest of the program...
Arabic sentence	...كان ضيف الحلقة الشيخ محمد حسان...
Buckwalter transliteration	...kAn Dyf AlHlqp Al$yx mHmd HsAn ...

3.2.1 English to French NE Transliteration

The proposed method is based on some linguistics resources, such as an English NE list and an English-French parallel corpus. It exploits the fact that English and French languages use similar alphabets and sound systems. This makes the transliteration process from English to French easier than between two languages that are completely distant, such as Arabic and French. Algorithm 1 describes formally the process of the English to French NE transliteration.

Thus, given an English NE as $EngNE$, we look for an English sentence $Seng(i)$ in the English-French parallel corpus that contains the $EngNE$ where i is the index of the English sentence in the English-French parallel corpus. Next, we tokenize the French sentence $Sfr(i)$ and we calculate the phonetic similarity between $EngNE$ and each word $Wfr(j)$ in the French sentence $Sfr(i)$. The pair ($EngNE$, $Wfr(j)$) that has the highest similarity score is considered as best transliteration noted $FrNE$. Tokenization of French sentences includes simple separation of words from punctuation marks, excluding the word "aujourd'hui" that is considered in French as a single word.

Algorithm 1 : *English to French NE transliteration*

```
Input: Parallel English-French corpus, EngNE.
Output: FrNE.
Seng(i)=Sentence containing EngNE;
Best_sim=0;
FrNE=Wfr(0);
For(each Wfr(j) in Sfr(i))
{
    Sim(j)=Similarity(EngNE, Wfr(j))
    if(Sim(j) > Best_sim)
    {
        Best_sim=Sim(j);
        FrNE=Wfr(j);
    }
}
```

3.2.2 Transliteration Similarity

To measure the transliteration similarity between an English NE and a French word, we use the Editex technique [25], based on a variant of Levenshtein edit distance algorithm [15] (kq, dt, lr, mn, gj, fpv, sxz, csz); such letters in a similar group frequently correspond to a similar pronunciation.

As in Levenshtein distance, the minimal number of insertions, deletions, and replacements necessary to transform one string to another is computed. But edits that replace a letter with another letter from a different group are weighted more heavily, and deletions of letters that are frequently silent (h and w) are weighted less heavily than other deletions.

ISSN 2007-9737

4 Evaluation

This section reports the evaluation of the Arabic-French NE translation system as well as the data used in the experiments. We evaluate our method on person and location NE types since other types, such as organization and date, are usually translated or a combination of transliteration and translation is used.

To evaluate the effectiveness of our new approach for Arabic-French NE recognition and translation, first, we describe the data used in our experiments. Second, we present the evaluation of the Arabic NE recognition task. Then, we measure the impact of the Arabic-French NE translation pairs on the performance of an in-house machine translation system, with a participation in TRAD2014 evaluation campaign[5].

4.1 The Data

The data used for the process of English-Arabic NE extraction consists of the United Nation (UN) English-Arabic parallel corpus and a parallel corpus extracted from Wikipedia. The size of the UN corpus used for the experiment is 377.7 M sentence pairs. The corpus consists of UN documents published on the web. The corpus of Wikipedia consists of bilingual titles related by an inter-language link of the source and target language. We extracted 13.8 M pairs of English-Arabic titles from the English and Arabic Wikipedia dump. For the process of English-French NE extraction, we used the English-French UN parallel corpus aligned on the sentence level and an English-French parallel corpus extracted from Wikipedia. The UN corpus consists of the United Nation parallel corpus freely available on the OPUS website[6]. This corpus consists of 13.2 M sentences. The corpus of titles consists of 13.8 M English-French phrases extracted from the English and the French Wikipedia. We notice that the UN corpus is characterized by its long sentences whereas the Wikipedia corpus contains short fragments.

[5] http://www.trad-campaign.org/

[6] http://opus.lingfil.uu.se/MultiUN.php

4.2 Evaluation of the Arabic NE Recognition Task

We have tested the Arabic NE discovery task with a randomly chosen 3,882,645 sentences from the English-Arabic UN corpus and 13,000 English-Arabic Wikipedia titles.

We have used three evaluation measures:

$$Precision = \frac{number\ of\ correctly\ transliterated\ NE}{number\ of\ transliterated\ NE},$$

$$Recall = \frac{number\ of\ correctly\ transliterated\ NE}{number\ of\ NE\ in\ the\ corpus},$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}.$$

Table 5 shows the precision, recall, and F-measure of our proposed method for the Arabic NE recognition task. These values were calculated based on the Arabic side of the Arabic-English lexicon. The precision values for the person and location types indicate that nearly all mined Arabic NEs are correct. The F-measure values exceed 78% for the two types. The results are very favorable and demonstrate the effectiveness of the Arabic NE detection task.

4.3 Evaluation of the Arabic-French NE Translation

We extracted 7,305 Arabic-French pairs of NE translations of the person name type and 4,250 of the location type. We evaluated the effectiveness of the overall framework using the precision measure by comparing translations produced by our system with those produced by Google Translate[7]. We translate the Arabic side of our lexicon using the Arabic to French Google machine translation and we compare our translation with that produced by Google. Table 6 presents the results in terms of the precision measure. We clearly observed that our system outperforms Google Translate in translating person names; however, Google Translate gives better translations of the location type.

[7] https://translate.google.fr in December 2014.

Table 5. Arabic NE recognition accuracy

		Precision	Recall	F-measure
UN Corpus	Person	93.30	87.34	90.22
	Location	99.37	65.36	78.85
Wikipedia	Person	99.81	89.64	94.45
	Location	99.92	78.67	88.03

Table 6. Arabic-French NE translation precision

	Person	Location
Our system	95.4	94.3
Google Translate	90.7	98.2

4.4 Improving Machine Translation Quality with NE Translation

This section presents an extrinsic evaluation of our approach of Arabic into French translation. We performed experiments on the Arabic to French machine translation. The Arabic-French machine translation is a phrase-based statistical machine translation system. It relies on two major components: phrase translation models and a DP-based phrase decoder [12]. The phrase translation pairs are extracted via word alignment, projection, and extension algorithms. Our SMT system participated in the second edition of TRAD[8] evaluation campaign.

In order to demonstrate the performance of our NE lexicons, we introduced two SMT systems. The first one is the baseline system that incorporates only the data offered by the TRAD campaign, whereas the second system (+NE) adds our developed Arabic-French NE lexicon to the TRAD data. The translation model of the baseline system was trained on the news-commentary, multiUN, and nist08 corpora. The language model was trained on the target side of these parallel data and the French side of the Europarl corpus. The development data is the test data of the first edition of TRAD (2012). The test data of our system is the test data supplied by the second edition of TRAD (2014). It is composed of 352 Arabic sentences with two references.

[8] http://www.trad-campaign.org/

Table 7. The size of the Arabic-French NE translation lexicons

	Geoname	JRC	Wikipedia	UN
Person	-	89	6806	450
Location	1829	-	5318	732
Total	15224			

Table 8. Improvement of SMT quality after introducing the Arabic-French NE translation component

	BLEU	TER	OOV
Baseline	29.0	59.8	2.99
+NE	29.7	59.4	2.63

The Arabic-French JRC NE list [21] and a list of NEs extracted from the Geoname[9] are also added to our lexicon. Table 7 shows the size of the data.

Gahbiche-Braham et al. [8] explored several strategies for the integration of NE lexicons in a SMT system. They demonstrated that the strategy of adding an NE lexicon to the baseline parallel data to learn one translation model is the best one for the location and the person types. We adopted such strategy and we added our NE translation lexicon to the training data of the baseline system and tested them against the test data. Table 8 illustrates the machine translation results in terms of BLEU [18], TER [18], and rates of OOVs. The BLEU score uses a modified form of precision to compare a candidate translation against multiple reference translations. TER (Translation Error Rate) is an error metric for machine translation that measures the number of edits required to change a system output into one of the references. OOV (Out Of Vocabulary) word rate is the rate of words that have not been translated by the machine translation system.

[9] http://www.geonames.org

The BLEU and TER values of our results are calculated by the TRAD campaign organizers. As shown in Table 8, "Baseline" is the system trained on the parallel corpora supplied by the TRAD campaign and "+NE" is the system trained on the baseline parallel data and NE translation lexicon.

When adding our NE translation lexicon, an obvious improvement could be observed in the BLEU score as well as in the TER score. This can be explained by some Arabic NEs in the test corpus that were messily translated into French in the baseline system. These NEs are correctly translated when we introduce the NE translation lexicon. Also, the ratio of OOV words decreases when adding the NE lexicon to the machine translation training data.

5 Conclusion

NE recognition and translation is a very problematic task for most NLP applications, such as machine translation. This module plays an important role in boosting the performance of the phrase-based machine translation system. Moreover, NE translation is a challenging problem especially for less-resourced language pairs such as Arabic-French.

In this paper, we have proposed a cross-linguistic method for the recognition of Arabic NEs and their translation/transliteration into French using a well-resourced language (English) as pivot. First, an English-Arabic and an English-French NE lexicons were extracted from an English-Arabic and an English-French parallel corpus, respectively. Second, we merged the terms from the two bilingual lexicons using the pivot English language and some transliteration rules. The extracted Arabic-French NE translation pairs were evaluated in a first step in terms of precision, recall, and F-measure and in a second step using an in-house statistical machine translation.

Our participation in TRAD 2014 evaluation campaign for the Arabic-French pair of languages showed an improvement of the performance of the phrase-based statistical machine translation system after the integration of the Arabic-French NE lexicon.

In the future, we are interested by including other types of NEs such as organization names and dates, in order to extend the coverage of the Arabic-French NE lexicon. Furthermore, comparable corpora will be exploited in addition to the parallel corpora in the NE translation/transliteration process and the phrase-based machine translation system.

References

1. **Al-Onaizan, Y. & Knight, K.** (**2002**). Translating named entities using monolingual and bilingual resources. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 400–408.

2. **Azab, M., Bouamor, H., Mohit, B., & Oflazer, K.** (**2013**). Dudley north visits north london: Learning when to transliterate to Arabic. *Proceedings of HLT/NAACL 2013, short papers section*, Atlanta, USA, pp. 439–444.

3. **Buckwalter, T.** (**2002**). *Buckwalter Arabic Morphological Analyzer Version 1.0.* Linguistic Data Consortium, University of Pennsylvania. Catalog: LDC2002L49.

4. **Chen, H.-H., Yang, C., & Lin, Y.** (**2003**). Learning formulation and transformation rules for multilingual named entities. *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition - Volume 15*, MultiNER '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–8.

5. **Darwish, K.** (**2010**). Transliteration mining with phonetic conflation and iterative training. *Proceedings of the 2010 Named Entities Workshop*, NEWS '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 53–56.

6. **Feng, D., Lü, Y., & Zhou, M.** A new approach for English-Chinese named entity alignment. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing EMNLP*.

7. **Finkel, J. R., Grenager, T., & Manning, C.** (**2005**). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370.

8. **Gahbiche-Braham, S., Bonneau-Maynard, H., & Yvon, F.** (**2014**). Traitement automatique des entites nommees en arabe : detection et traduction. *TAL (Traitement Automatique des Langues)*, Vol. 5, No. 2.

9. **Gupta, P., Rao, S., & Majumder, P.** (**2010**). External plagiarism detection: N-gram approach using named entity recognizer - lab report for PAN at CLEF 2010. *CLEF (Notebook Papers/LABs/Workshops)*.

10. **Habash, N.** (**2008**). Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 57–60.

11. **Huang, F., Vogel, S., & Waibel, A.** (**2003**). Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition - Volume 15*, MultiNER '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 9–16.

12. **Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E.** (**2007**). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 177–180.

13. **Kumano, T., Kashioka, H., Tanaka, H., & Fukusima, T.** (**2004**). Acquiring bilingual named entity translations from content aligned corpora. *IJCNLP*, pp. 177–186.

14. **Lee, C.-J., Chang, J. S., & Jang, J.-S. R.** (**2006**). Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 5, No. 2, pp. 121–145.

15. **Levenshtein, V. I.** (**1966**). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*, pp. 707–710.

16. **Liu, J. & Birnbaum, L.** (**2008**). What do they think?: Aggregating local views about news events and topics. *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, ACM, New York, NY, USA, pp. 1021–1022.

17. **Moore, R. C.** (**2003**). Learning translations of named-entity phrases from parallel corpora. *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 259–266.

18. **Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J.** (**2002**). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 311–318.

19. **Samy, D., Moreno, A., & Guirao, J. M.** (**2005**). A proposal for an Arabic named entity tagger leveraging a parallel corpus (Spanish-Arabic). *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pp. 459–465.

20. **Sellami, R., Sadat, F., & Belguith Hadrich, L.** (**2014**). Mining named entity translation from non parallel corpora. *FLAIRS Conference*, pp. 219–224.

21. **Steinberger, R., Pouliquen, B., Kabadjov, M. A., & der Goot, E. V.** (**2013**). JRC-Names: A freely available, highly multilingual named entity resource. *CoRR*, Vol. abs/1309.6162.

22. **Zaghouani, W.** (**2012**). RENAR: A rule-based arabic named entity recognition system. *ACM Trans. Asian Lang. Inf. Process.*, Vol. 11, No. 1, pp. 2–13.

23. **Zaghouani, W.** (**2014**). Critical survey of the freely available Arabic corpora. *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC*, pp. 1–8.

24. **Zitouni, I. & Florian, R.** (**2008**). Mention detection crossing the language barrier. *Proceedings of 2008 Conference on Empirical Methods in Natural Language Processing EMNLP*, pp. 600–609.

25. **Zobel, J. & Dart, P. W.** (**1996**). Phonetic string matching: Lessons from information retrieval. **Frei, H.-P., Harman, D., Schäuble, P., & Wilkinson, R.**, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, ACM, pp. 166–172.

ISSN 2007-9737

Rahma Sellami is a doctoral student in Computer Sciences at the University of Sfax, Tunisia. She is a researcher at ANLP Research Group of MIRACL Laboratory. Her Ph.D. thesis aims to exploit comparable corpora for statistical machine translation. Her main interest focuses on Arabic language processing.

Fatima Deffaf is a master student in Computer Sciences at the University of Quebec in Montreal, Canada. Her master thesis aims to exploit parallel corpora for named entity recognition and translation.

Fatiha Sadat is an Associate Professor at the University of Quebec in Montreal, Canada. She received her doctoral degree in 2003 from the Computer Science Department, Nara Institute of Science and Technology, Nara, Japan. Her research includes work on cross-language information retrieval, social media analysis, multilingual ontologies, machine translation, natural language processing, morphological analysis and computational analysis of Arabic dialects. In the past, Fatiha Sadat was a researcher at the National Research Council of Canada and the National Institute of Informatics, as a post-doctoral fellow under the JSPS program (Japan Society for the Promotion of Science).

Lamia Hadrich Belguith is a Professor of Computer Science at Sfax University, Tunisia, and Head of the Arabic NLP Research Group at MIRACL Laboratory. Her research interest is mainly focused on Arabic language processing and its applications. She is also interested in a number of other topics such as summarization, question answering, and Tunisian dialect processing. She has published extensively in her field.

Article received on 27/01/2015; accepted on 05/03/2015.
Corresponding author is Rahma Sellami.