Using Transliteration of Proper Names from Arabic to Latin Script to Improve English-Arabic Word Alignment

Nasredine Semmar
Institut CEA LIST, Laboratoire Vision et Ingénierie des Contenus
CEA Saclay – Nano-INNOV, F-91191 Gif-sur-Yvette Cedex, France
nasredine.semmar@cea.fr

Houda Saadane
LIDILEM, Université Grenoble III
Domaine Universitaire
1180, avenue centrale, F-38400 Saint Martin d'Hères, France
houda.saadane@e.u-grenoble3.fr

Abstract

Bilingual lexicons of proper names play a vital role in machine translation and cross-language information retrieval. Word alignment approaches are generally used to construct bilingual lexicons automatically from parallel corpora. Aligning proper names is a task particularly difficult when the source and target languages of the parallel corpus do not share a same written script. We present in this paper a system to transliterate automatically proper names from Arabic to Latin script, and a tool to align single and compound words from English-Arabic parallel texts. We particularly focus on the impact of using transliteration to improve the performance of the word alignment tool. We have evaluated the word alignment tool integrating transliteration of proper names from Arabic to Latin script using two methods: A manual evaluation of the alignment quality and an evaluation of the impact of this alignment on the translation quality by using the open source statistical machine translation system Moses. Experiments show that integrating transliteration of proper names into the alignment process improves the F-measure of word alignment from 72% to 81% and the translation BLEU score from 20.15% to 20.63%.

1 Introduction

Bilingual lexicons of proper names play a vital role in Machine Translation (MT) and Cross-Language Information Retrieval (CLIR). Word alignment approaches are generally used to construct bilingual lexicons automatically from parallel corpora. Aligning proper names requires both recognition of the proper names present in the parallel corpus and their alignment (Abuleil and Evens, 2004). This task is particularly difficult when the source and target languages of the parallel corpus do not share a same written script. A solution to this issue consists in writing the proper names present in the parallel corpus in the same written script. This operation is named transliteration and consists in replacing each grapheme of a writing system by another grapheme or a group of graphemes of another writing system, regardless of pronunciation.

In order to study the impact of using transliteration to improve the performance of a word alignment tool, we present in this paper a system to transliterate automatically proper names from Arabic to Latin script, and a tool to align single and compound words from English-Arabic parallel texts.

The remainder of the paper is organized as follows: Section 2 recalls in some previous work addressing tasks of transliteration and bilingual lexicon extraction from parallel corpora. In section 3, we present briefly the system for automatic transliteration of proper names from Arabic to Latin script. Section 4 describes the process of using transliteration in the word alignment tool. We present in section 5 the experimental protocol we followed and discuss the obtained results. We finally conclude and present directions for future work in section 6.

2 Related Work

In order to build bilingual lexicons from parallel corpora automatically, several word alignment approaches have been explored (Daille *et al*., 1994; Blank, 2000; Barbu, 2004). These approaches align proper names correctly when the source and target languages of the parallel corpus

International Joint Conference on Natural Language Processing, pages 1022–1026,
Nagoya, Japan, 14-18 October 2013.

share a same written script. Recent research works for aligning proper names when the source and target languages do not share a same written script have focused on automatic alignment of transliterations in order to enrich bilingual lexicons of named entities. These include (Al-Onaizan and Knight, 2002) and (Sherif and Kondrak, 2007) who worked on the Arabic-English alignment, (Tao *et al.*, 2006) who worked on Arabic, Chinese and English, and (Shao and Ng, 2004) who used the information resulted from transliterations based on pronunciation. They combined the obtained information from the translation context and those generated from Chinese and English transliteration. This technique allowed processing some specific infrequent words. Some other systems assign for a given name only one transliteration such as the generative model for English words written in Japanese (Katakana) to Latin transcription (Knight and Graehl, 1997). This approach was adapted by (Stalls and Knight, 1998) to translate English words written in Arabic into English. (AbdulJaleel and Larkey, 2003) proposed a system based on a statistical approach to transliterate English names into Arabic. This system has several limitations as it uses the computation of the most probable form supposed to be the correct one. Indeed, this hypothesis is not always valid in all the Arab countries and dialects. To avoid pronunciation and dialect varieties, (Alghamdi, 2005) proposed a system to transliterate vowelized Arabic names into English. This system is based on a dictionary of Arabic names in which the pronunciation is set using vowels added to listed names with an indication of their equivalents in English. This approach has a strong limitation when used in word alignment as it proposes only one transliteration for a given name. Recently, (Saadane *et al.*, 2012) proposed an approach to transliterate proper names from Arabic to Latin script which takes into account phonological and linguistic aspects. The authors reported an improvement of the F-measure of their French-Arabic word alignment tool from 82% to 86%.

3 Transliteration of Proper Names from Arabic to Latin Script

The transliteration system of proper names from Arabic to Latin script used in this study (Saadane *et al.*, 2012) is based on a finite-state automaton. This automaton switches from one state to another according to the outward transitions of the current state and the currently processed letter of the Arabic word. The transliteration process is composed of the following main steps:

1. Transliteration: Each proper name is, first, split or not into several elements according to its type and the particles which do not compose the name itself are transcribed. Then, transliteration rules are applied to transliterate the names themselves. These rules are applied in a certain order based on the number of consonants of the proper name. For example, the compound name "عبد الرشيد" is, first, split into "عبد + ال + رشيد", second, the particles "عبد" and "ال" are transcribed into "abd" and "al", and finally the name "رشيد" is transliterated into rachid, rashid, etc.
2. Normalization: This step consists in performing some post-processing on the generated transliterations such as changing the first letter into capital.
3. Weighting**:** This step consists in assigning weights to the rules used to generate the list of transliterations in order to display the results sorted from the most likely to the least likely. Results of some search engines are exploited to compute these weights based on the number of occurrences for each generated transliteration of the proper name.

4 Alignment of Proper Names from English-Arabic Corpora

Word alignment from parallel corpora consists, on the one hand, in identifying words present in the source and target texts, and, on the other hand, in establishing correspondences between these words. The word alignment tool evaluated in this study (Semmar *et al.*, 2010), first, identifies single words and compound words present in the parallel corpus using the linguistic analyzer LIMA (Besançon *et al.*, 2010), and, second, establishes correspondence relations between these words using the following steps:

1. Look-up of words which are present in an existing English-Arabic lexicon composed of 149495 entries;
2. Matching of words which are cognates;
3. Matching of words which have the same grammatical categories;
4. Establishing correspondence relations between compound words.

We describe below only the step 2 which illustrates the process of using transliteration of proper names from Arabic to Latin script in English-Arabic word alignment.

Proper names alignment consists, first, in searching words present in the source and target sentences which have the grammatical category "Proper Name" by using the results of the linguistic analyzer LIMA, and, second, in identifying words which are cognates. Several research works have shown that using cognates can improve both sentence alignment (Simard *et al.*, 1993) and word alignment (Kondrak, 2005). In our implementation, we consider, in a first step, that pairs of words which share the first four characters as cognates. This step uses the results of the transliteration into Latin script of all the proper names present in the Arabic corpus and can identify, for example, that the proper name "Kosovo" and the transliteration of the Arabic word "كوسوفو" ("kosoufou") are cognates. However, this step does not detect pairs of words such as "Algeria" and "aljazair" (transliteration of the Arabic word "الجزائر"). To take into account this kind of pairs of words, we used the Jaro–Winkler distance *DJW* (Winkler, 1990), a similarity measure based on the number of letters in common between the string of the word of the source language *ws* and the string of the word of the target language *wt*.

$$DJ(ws,wt) = \begin{cases} 0 & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|ws|} + \frac{m}{|wt|} + \frac{m-t}{m}\right) & otherwise \end{cases}$$

where:

- *m* is the number of matching characters. Two characters from *ws* and *wt* respectively, are considered matching only if they are the same and not farther than:

$$\left(\frac{max(|ws|,|wt|)}{2}\right) - 1$$

- *t* is the number of transpositions which is equal to the half of number of characters in *ws* that do not line up (by index in the matched subsequence) with identical characters in *wt*.
- |*ws*|, |*wt*| are lengths of the strings corresponding to the words *ws* and *wt*.

Jaro–Winkler similarity measure is a variant of the Jaro distance metric *DJ* (Jaro, 1989).

$$DJW(ws,wt) = DJ(ws,wt) + (lp(1 - DJ(ws,wt)))$$

where:

- *l* is the length of common prefix at the start of the string up to a maximum of 4 characters.
- *p* is a constant scaling factor for how much the score is adjusted upwards for having common prefixes.

In order to identify the values of *l* and *p* which provide the best alignment, we checked manually the result of the transliteration of 254 proper names. This evaluation showed that, if *l* is equal to 2 and *p* is equal to 0.1, the words *ws* and *wt* are cognates when the value of the Jaro–Winkler distance is the highest. Table 1 presents results after running our word alignment tool on the English sentence "Condemning all violations of human rights in Kosovo which have affected all ethnic groups in Kosovo." and its Arabic translation "وإذ تدين كل ما ارتكب في كوسوفو من انتهاكات لحقوق الإنسان طالت جميع الفئات العرقية في كوسوفو.".

Lemmas of words of the source sentence	**Lemmas of words of the target sentence**
condemn	أَدَانَ
violation	إِنْتِهَاك
human	إِنْسَان
right	حَق
Kosovo	كوسوفو
affect	طَالَ
ethnic	عَرَقِيَّة
group	فِئَة
Kosovo	كوسوفو
violation_human_right	إِنْتِهَاك_حَق_إِنْسَان
human_right	حَق_إِنْسَان
ethnic_group	فِئَة_عَرَقِيَّة

Table 1. Results of single and compound words alignment

The word "Kosovo" was aligned using cognates matching after transliteration, the words "condemn", "human", "affect" and "group" were aligned using grammatical categories matching and the other single words exist in the English-Arabic lexicon. The compound words "violation_human_right", "إِنْتِهَاك_حَق_إِنْسَان", "human_right", "حَق_إِنْسَان", "ethnic_group" and "فِئَة_عَرَقِيَّة" are first recognized by LIMA respectively from the source sentence and the target sentence, and then aligned using lexical and syntactic transfer rules between source and target languages (Ozdowska, 2004).

5 Experimental Results and Evaluation

The impact of using transliteration of proper names on the quality of alignment and machine translation has been evaluated according to the two following approaches:

- A manual evaluation comparing the results of our word aligner with a reference alignment;
- An automatic evaluation by integrating the results of our word aligner tool in the training corpus used to build the translation table of the statistical MT system Moses (Koehn *et al.*, 2007).

In order to evaluate the alignment quality manually, we used 500 English-Arabic aligned sentences extracted from the MT evaluation MEDAR[1] package and we followed the evaluation framework defined in (Mihalcea and Pedersen, 2003). Table 2 summarizes the results of our word aligner in terms of precision and recall. The first line describes the performance of the word aligner when it does not integrate transliteration and the second line mentions its performance when it uses transliteration. As we can see, the results demonstrate that using transliteration improves both precision and recall of word alignment. These results confirm those obtained by (Sajjad *et al.*, 2003) related to the improvement of alignment quality when integrating transliteration into the GIZA++ word aligner.

Alignment	**Precision**	**Recall**	**F-measure**
without using transliteration	0.90	0.60	0.72
with the use of transliteration	0.91	0.73	0.81

Table 2. Results of the evaluation of single and compound words alignment

The unavailability of a reference alignment of a significant size for single and compound words does not allow us to compare our approach with the state-of-the-art work. That's why we decided to study the impact of the use of transliteration in word alignment by integrating the results of our word aligner in the training corpus used to extract the translation model of Moses. The initial training corpus is composed of 75000 pairs of English-Arabic sentences extracted from the MEDAR corpus (2631654 English words and 2344878 Arabic words). We added to this corpus around 28000 pairs of single and compound words corresponding to the results of our word aligner which integrates transliteration applied on 1000 pairs of English-Arabic sentences. We also specified a language model for the target language using a corpus composed of 100000 Arabic sentences (3155516 words). The performance of the statistical machine translation system Moses is evaluated using the BLEU score on a test corpus composed of 500 pairs of sentences. Note that we consider one reference per sentence. The obtained results show that the inclusion in the training corpus of word alignment results integrating transliteration has improved the translation BLEU score from 20.15 to 20.63 (a gain of 0.48 points).

In order to assess statistical significance of the obtained results, we use the paired bootstrap resampling method (Koehn, 2004) which estimates the probability (*p-value*) that a measured difference in BLEU scores arose by chance by repeatedly (10 times) creating new virtual test sets by drawing sentences with replacement from a given collection of translated sentences. We carry out experiments using this method to compare the translation results without using transliteration and with the use of transliteration. At a 95% confidence interval (CI), the results vary from insignificant (at $p > 0.05$) to highly significant. The *p-value* obtained is equal to 0.02 and therefore the improvement achieved by using transliteration is statistically significant.

6 Conclusion

We presented briefly in this paper a system to transliterate proper names from Arabic to Latin script and we proposed a tool to automatically align word pairs from an English-Arabic parallel corpus. We integrated the transliterated proper names into the cognates matching step and we obtained a gain of 9% on word alignment F-measure and a gain of 0.48 points in translation BLEU score. These encouraging results can be improved in a number of ways. First, we plan to affect a weight for each word pair in order to filter the word alignment results and to integrate them directly in the translation table of Moses. We also plan to use, on the one hand, the linguistic analyzer LIMA to lemmatize texts of the bilingual corpus, and on the other hand, factored models and a flexor to generate adequate surface forms from lemmas.

[1] The MT evaluation MEDAR package is available on http://www.medar.info/index.php.

References

Saleem Abuleil and Martha Evens. 2004. *Named Entity Recognition and Classification for Text in Arabic*. The 13th International Conference on Intelligent & Adaptive Systems and Software Engineering, Nice, France.

Nasreen AbdulJaleel and Leah S. Larkey. 2003. *Statistical transliteration for English-Arabic Cross Language Information Retrieval*. The 12th ACM International Conference on Information and Knowledge Management, New Orleans, LA, USA.

Mansour Alghamdi. 2005. *Algorithms for Romanizing Arabic names*. Journal of King Saud University. Computer Sciences and Information. Riyadh, 17: Pages 1-27.

Yaser Al-Onaizan and Kevin Knight. 2002. *Translating named entities using monolingual and bilingual resources*. The 40th ACL Conference, Philadelphia, USA.

Ana M. Barbu. 2004. *Simple linguistic methods for improving a word alignment algorithm*. The 7th International Conference on the Statistical Analysis of Textual Data (JADT), Louvain, Belgium.

Romaric Besançon, Gaël De Chalendar, Olivier Ferret, Faïza Gara, Meriama Laib, Olivier Mesnard, and Nasredine Semmar. 2010. *LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation*. The 7th international conference on Language Resources and Evaluation, Valletta, Malta.

Ingeborg Blank. *Terminology extraction from parallel technical texts*. Véronis J. (Ed.), Parallel Text Processing, Dordrecht: Kluwer, 2000.

Béatrice Daille, Eric Gaussier, and Jean. M. Langé. 1994. *Towards automatic extraction of monolingual and bilingual terminology*. The 15th International Conference on Computational Linguistics.

Matthew A. Jaro. 1989. *Advances in record linkage methodology as applied to the 1985 census of Tampa Florida*. Journal of the American Statistical Association 84: Pages 414–420.

Kevin Knight and Jonathan Graehl. 1997. *Machine transliteration*. Journal Computational Linguistics, 24(4): Pages 599-612.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicolas Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: Open source toolkit for statistical machine translation*. The Conference ACL 2007, demo session, Prague, Czech Republic.

Philipp Koehn. 2004. *Statistical significance tests for machine translation evaluation*. The 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.

Grzegorz Kondrak. 2005. *Cognates and Word Alignment in Bitexts*. The Tenth Machine Translation Summit (MT Summit X), Phuket, Thailand.

Rada Mihalcea and Ted Pedersen. 2003. *An evaluation exercise for word alignment*. The Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, Canada.

Sylwia Ozdowska. 2004. *Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora*. The 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Houda Saadane, Nasredine Semmar, Ouafa Benterki and Christian Fluhr. 2012. *Using Arabic Transliteration to Improve Word Alignment from French-Arabic Parallel Corpora*. The fourth Workshop on Computational Approaches to Arabic Script-based Languages, AMTA 2012, San Diego, CA, USA.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2011. *An algorithm for unsupervised transliteration mining with an application to word alignment*. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Pages 430-439, Portland, Oregon, USA

Li Shao and Hwee Tou Ng. 2004. *Mining new word translations from comparable corpora*. The 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland.

Tarek Sherif and Grzegorz Kondrak. 2007. *Bootstrapping a stochastic transducer for Arabic-English transliteration extraction*. The 45th ACL Conference, Prague, Czech Republic.

Nasredine Semmar, Christophe Servan, Gaël de Chalendar, Benoît Le Ny, and Jean-Jacques Bouzaglou. 2010. *A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions*. The 32nd Translating and the Computer Conference, England.

Michel Simard, George F. Foster, and Pierre Isabelle. *Using cognates to align sentences in bilingual corpora*. 1993. The Conference of the Centre for Advanced Studies on Collaborative Research: Distributed computing, Volume 2, Pages 1071-1082.

Bonnie Stalls and Kevin Knight. 1998. *Translating names and technical terms in Arabic text*. The COLING/ACL Workshop on Computational Approaches to Semitic Languages, Montreal, Canada.

Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat, and ChengXiang Zhai. 2006. *Unsupervised named entity transliteration using temporal and phonetic correlation*. The 2006 EMNLP Conference, Pages 250-257.

William E. Winkler. 1990. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Section on Survey Research Methods, American Statistical Association: Pages 354–359.