Using Arabic Transliteration to Improve Word Alignment from French-Arabic Parallel Corpora

Houda Saadane

LIDILEM - Université Stendhal Grenoble 3
BP 25, 38040 Grenoble Cedex, France
houda.saadane@e.u-grenoble3.fr

Ouafa Benterki, Nasredine Semmar, Christian Fluhr

Institut Supérieur Arabe de Traduction
Rue Tabrizi, 16013 Bir Mourad Raïs, Algérie
obenterki@hotmail.com, nasredine.semmarn@club-internet.fr
christian.fluhr@gmail.com

Abstract

In this paper, we focus on the use of Arabic transliteration to improve the results of a linguistics-based word alignment approach from parallel text corpora. This approach uses, on the one hand, a bilingual lexicon, named entities, cognates and grammatical tags to align single words, and on the other hand, syntactic dependency relations to align compound words. We have evaluated the word aligner integrating Arabic transliteration 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 Moses statistical machine translation system. The obtained results show that Arabic transliteration improves the quality of both alignment and translation.

1 Introduction

Transcription consists in replacing each sound or phoneme of a phonological system by a grapheme or a group of graphemes of a writing system, while transliteration consists in replacing each grapheme of a writing system by another grapheme of a group of graphemes of another writing system, regardless of pronunciation. The objective transcription is to reconstruct the original pronunciation using the writing system of the target language and the goal of the transliteration is to represent the original grapheme with the corresponding graphemes of the target languages.

Transcription and transliteration are experiencing significant growth due to the increasingly multilingual Internet and to the exponential needs in the field of cross-language information retrieval (CLIR). This is especially true for finding named entities (names of persons, places, companies, organizations, etc.) but these entities have a plurality of forms, spellings, and transcripts depending on languages and countries. The case of Arabic names illustrates this complex and multifaceted situation. For example, the name of the Libyan leader (Gaddafi), which has a single spelling in Arabic (معمر القذافي) but several pronunciations and accents depending on the dialect, is transcribed into Latin script by over 60 different forms, including: Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi,Muammar Gadafi, Moamer El Kazzafi, Mu'ammar al-Qadhdhafi, Mu'amar Qadafi, Muammar Gheddafi, Mu'ammar Al Qathafi, Muʿammar Al-Qaḍâfî…

In this paper, we first outline the theoretical issues and practical difficulties that arise in the transliteration of names and surnames and possible treatments that could resolve these difficulties.

Then, we present, on the one hand, our system for automatic transliteration of Arabic names, and on the other hand, the impact of using transliteration to improve the performance of a word alignment tool.

2 Related Work

The transliteration problem has interested many linguists in different languages, and recently researchers in natural language processing due to the constant development and use of Internet. Many research works have focused on the automatic alignment of transliterations from a multilingual text corpus, in order to enrich bilingual lexicons, which play a vital role in machine translation (MT) and cross-language information retrieval. These include (Al-Onaizan and Knight, 2002) and (Sherif and Kondrak, 2007) who worked on the Arabic-English alignment, (Tao et al., 2006) who work on Arabic, Chinese and English and (Shao and Ng, 2004) who use the information resulted from transliterations based on pronunciation. (Shao and Ng, 2004) combine the obtained information from the translation context and those generated from the Chinese and English transliteration. This technique allows processing some specific infrequent words. We can also find some other systems that 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 an English word written in Arabic to English. The system of transliteration generation is based on a training dictionary that considers the unknown and unlisted pronunciations within the system. In order to resolve this deficiency, some works have used statistical techniques. This is the case of the transliteration system of the English names to Arabic proposed by (AbdulJaleel and Larkey, 2003). However, this system has several limitations as it uses the computation of the most probable form, supposed to be the correct form but is not always valid in all the Arab countries and dialects. To avoid the pronunciation and dialect's flavor problems, (Alghamdi, 2005) has proposed a transliteration system to translate vowelized Arabic names written in 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. Meanwhile, this approach cumulates the disadvantages of the previous techniques: it does considerate the unlisted pronunciations in the dictionary and it is normative as it proposes only one transliteration for a given name. Apparently, the author favored the adoption of a standard transliteration, but this can be only a personal isolated initiative.

Globally, the current works on transcription and transliteration do not reflect their complexity that affects both the oral and writing in two or more linguistic systems in the same time. In fact, transcribing a name from a source linguistic system to another target system is a delicate task which needs some operations requiring management of a set of morphologic, phonetic and semantic properties. These operations are necessary to ensure a robust transliteration process, especially for security, checking identity or information retrieval applications.

However, few studies consider the links between:

- compared phonology and inter-lingual transcription,
- compared graphematic and inter-lingual transcription,
- Arabic dialectology and Latin transliteration systems.

The few studies propose a solution treating partially one of these problematics dedicated to the automatic identification of the speaker origin from its dialect. It is the case of the mentioned studies in (Guidère, 2004) and (Barkat-Defradas et al., 2004).

3 Transliteration of Names Written in Standard Arabic in Latin Characters

The Arabic transcription system includes 28 letters: 25 consonants and 3 vowels that can be short or long according to the word. It contains also some specific morphological and phonological phenomena that must to be taken into account in a transliteration process as the duplication of consonants, sometimes materialized in the Arabic transcription by "shadda", and the repetition of vowels referenced in Arabic by "tanwin". But the

modern Arabic transcription presents the particularity of omitting in general from the texts the indications to the vowels repetition or the short vowels which constitute a source of ambiguity for the transliteration systems.

3.1 Methodology of the Transliterator Construction

We have chosen a "bottom-up" methodology to construct our transliterator. We first start by identifying the existing transliterations for each Arabic letter from the usage norms observed on Internet. This empiric investigation is based on a corpus of texts collected in different languages targeted by the transliterator. It allows to construct a library of graphematic equivalences currently used in the texts transcribed in Latin. In the following table, we present some graphematic equivalences extracted from the used corpus:

Arabic letter	Equivalences in Latin
ء	a
ا	A, a, ä, â, á, ā, e, ê
ب	B, b
ت	T, t
ث	Th, th, t, ṯ
غ	Gh, gh, Ğ, ğ, ḡ
ف	F, f, ph
ق	Q, q, C, c, K, k
ك	K, k, C, c
ل	L, l

Table 1: Some graphematic equivalences between Arabic and Latin alphabets.

The study of the corpus allows us to observe that some Arabic letters, without graphematic equivalence in Latin transcription, was transcribed by some Arabic digits in the text written in Latin. This kind of transliteration is particularly used in phone messages (SMS) and the social networks in Europe or Middle East. The following table summarizes these alphanumeric equivalences for the concerned Arabic letters:

Arabic letter	Representation as a number
ء	2
ح	7
خ	7'
ص	9
ض	9'
ط	6
ظ	6'
ع	3
غ	3'
ق	8

Table 2: Transliterations of Arabic letters into numbers.

Hence, by combining these two types of symbolic representations, we can find in the translated texts these equivalences for the usual Arabic names:

Name in Arabic	منى	عدنان	حنان	طارق
Examples of equivalent transcriptions in Latin	Mouna or Mona...	Adnane or 3adnan...	Hanane or 7anan...	Tarek or 6ariq...

Table 3: Examples of Arabic names.

This variation in the use of transliteration is a source of ambiguity when we search information automatically. We can explain this phenomena as follows:

First, for some historical reasons Arab countries were colonized or remanded by some European countries during some periods which were different from one country to another. This occupation has affected the pronunciation, the vocabulary and the transliteration of names of the country's population. Thereby, the influence of the French graphematic and linguistic system is perceptible in the usages of the transliterations in the Maghreb countries, with different intensity from a country to another one. We can see the same thing in the Middle East countries with the English and American influences. Therefore, for political reasons, a common norm does not exist or a unified strategy in the field of transliteration for the Arabic language. This has led every writer or transcriber to use the most dialectal pronunciation

to transcribe the Arabic names. The famous example is that of Laurence of Arabia who, for transcribing the name of the Djeddah city in Saudi Arabia, (جدة), uses 25 times the spelling “Jeddah”, 6 times the spelling “Jidda” and one time the spelling “Jedda” in the same book (in 1926). Laurence of Arabia justified this variation in the transliteration by the following: «we cannot transcribe correctly and with the same manner an Arabic name because the differences between the Arabic and Latin consonants; and the vowel pronunciation which is different from a region to another one» (Alsalman et al., 2007). This is still always true as the different transcriptions of the “Jaddah” cited in Laurence of Arabia are actually used.

Finally, for dialectology reasons, it exists a variety of regional and local dialects in the Arab world. This variety renders impossible finding the same pronunciation for a set of regions or countries. For instance, one of the most frequently used names is the name of the prophet Muhammad (محمد). This name is transcribed in French by Mahomet and has many different pronunciations (transcriptions) like: {Mohamed, Mouhammad, Muhamed, Mhamed, M’Hamed, Muhammad...}. Even when the name is vowelized, it presents many possibilities of transliteration in the texts: {Muhamad, Mouhamad, Mohamad, Mehammad, Mehammade}.

This variation of transliteration according to the dialects is sometimes associated to the use of special characters in some Arab countries or regions. For instance, the following names represent some unconventional forms in Latin transcription: Mu`ammar, Mabrūk, Muṣṭafá, Ismā`īl, Hâdî. All these phenomena require an accurate observation during the process in order to identify the problems and construct efficient rules allowing an automatic process of Arabic names transliteration in real time.

3.2 Description of the Arabic to Latin Translitertor

The module of transliteration of the Arabic script to Latin script is based on finite-state machines (finite-state automata): it consists of states and conditional transitions. Its operation is determined by the nature of the input word: the 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. After processing its entire letters, the automaton accepts or rejects the input word. Then the vowels of the input word are removed (if any), and the transliteration is carried out. Finally, the module outputs a sorted list of Arabic names written in Latin characters.

The core of the transliteration system consists of contextual rules. These rules are intended to accurately model the observed forms in the input: is it a "kunya"? A name preceded by an article? Or a first name only?

According to (Guidère, 2006), the name of a person contains several elements in Arabic script. It consists in principle of four main components:

1. The "Kunya" (Particle): typically composed of "Abu" (father of) followed by a name of a child, or of "Umm" (mother of) followed by a name of a child. Example: "Abu Omar" (Father of Omar), "Umm Mohammed" (Mother of Mohammed),
2. The "Ism" (Name): for example, Omar, Ali Mohamed, Khaled Abdallah, etc. It indicates the ethnic or sectarian of the wearer: for example, "Omar" is a typically Sunni name, "Rustam" is a typically Iranian name,"Arslan" is typically Turkish, etc.
3. The "Nasab" (Genealogical affiliation): each name is preceded by "bin" or "Bin/Ben" (Bint/Bent for women). It indicates the exact genealogical descent of the underlying individual. Arabs sometimes go back very far in the indication of the ancestors to avoid confusion among people: ex. Muhammad Salih Bin Abdullah Bin Said Bin, etc.
4. The "Nisba" (suffix of origin): this suffix mainly refers in principle to the tribe or clan in the old genealogy but today it refers specifically to the birthplace of individuals: Maghribi (born in Morocco), Libi (born in Libya), Masri (born in Egypt), Djazaiiri (born in Algeria)…etc. The "Nisba" is always preceded by the

article [Al] and ends with the suffix [i]. It indicates the initial territorial residence of persons, or their nationality.

First, the particles, the part which is not the name itself, are transcribed. Then the transliteration rules are applied to transliterate the names themselves. These transliteration rules are applied in a certain order based on the number of consonants of the name in question and on priority weights. For example, let's consider the name "عبد (Abd) + AL (ال) + Name (رحمن)", the system proceeds as follows:

- Transliteration of the particle عبد (Abd);
- Transliteration of the article ال (Al);
- Concatenating the particle "Abd" and the article "Al" (with a space) and linking them to the name with a hyphen: Abd Al-Rahman (عبد الرحمن);
- Generation of all possible forms of transliteration for these three elements:

Arabic proper name	Transliterations
عبد الرحمن	Abd Al Rahman Abd al-Rahman Abd al Rahman Abd El-Rahman Abd El Rahman Abd el-Rahman Abd el Rahman Abd Ar-Rahman Abd Ar Rahman Abd Ar-Rahman Abd ar-Rahman

Table 4: Some transliteration forms for عبد الرحمن.

An intermediate step allows to overcome some of the very difficult problems of transcription, such as transcription of certain names whose pronunciations change completely for religious or other reasons: this is the case of Moussa translated into Moses, Yusuf into Joseph, Yaakoub into Jacob, Hawa into Eve, etc.

Once the sorted list of transliterated names is generated, the next two tasks are performed:

- Normalization of the list of names in Latin script: This step is to perform some post-processing on the output name in Latin script such as the removal of special characters (diacritics and figures) and changing the first letter into capital (capitalization does not exist in the Arabic script). This notion of capital is retained only in the case of use in databases, but it is not added to the usual search engines, which do not consider the case as relevant;
- Weighting of the output names in Latin script: This step consists in assigning a weight to the rules that were used to generate the list, in order to display the output results sorted from the most likely to the least likely, or vice versa. To achieve this weighting, we use various search engines and the number of occurrences for each generated form of the name: for example, for the Arabic name جمال (jamal), the system generates three different transliterations (Djamel, Jamel, Gamel) and search results frequencies give the following ratios:

Latin transliterated form of the name	Number of occurrences of the name
Djamel	4000000
Jamel	5500000
Gamel	500000

Table 5: Results with Google for the transliterated form of the name جمال.

From the perspective of weighting, this example shows that the Arabic letter (ج) is transcribed, in terms of frequency, mainly by (J), followed by (Dj) and finally by (G).

This procedure has been applied to all the forms of the transliteration of the Arabic characters. It allows establishing a weighted list of equivalences of graphemes that will be used to display the results from the most likely to the least one or vice versa.

4 Using Transliteration to Improve Word Alignment

Word alignment consists of finding correspondences between single words and compound words in a bilingual corpus aligned at

the sentence level. Our word alignment tool uses an existing bilingual lexicon and the following linguistic properties:

- named entities, positions and grammatical categories to align single words,
- syntactic dependency relations to align compound words.

These properties are produced by a linguistic analyzer which is built using a traditional architecture involving separate processing modules:

- A Tokenizer which separates the input text into a list of words.
- A Morphological analyzer which looks up each word in a general full form dictionary. If these words are found, they are associated with their lemmas and all their grammatical tags. For Arabic agglutinated words which are not in the full form dictionary, a clitic stemmer (Larkey et al., 2002) was added to the morphological analyzer. The role of this stemmer is to split agglutinated words into proclitics, simple forms and enclitics.
- A Part-Of-Speech (POS) tagger which searches valid paths through all the possible tags paths using attested trigrams and bigrams sequences. The trigram and bigram sequences are generated from a manually annotated training corpus.
- A Syntactic analyzer which is used to split the list of words into nominal and verbal chain and recognize dependency relations by using a set of syntactic rules. We developed a set of dependency relations to link nouns to other nouns, a noun with a proper noun, a proper noun with the post nominal adjective and a noun with a post nominal adjective. These relations are restricted to the same nominal chain and are used to compute compound words.
- A Named Entity recognizer which uses name triggers such as “Doctor”, “President”, “Government”... to identify named entities (Abuleil and Evens, 2004).

Word alignment using the existing bilingual lexicon consists in extracting for each word of the source sentence the appropriate translation in the bilingual lexicon. The result of this step is a list of lemmas of source words for which one or more translations were found in the bilingual lexicon.

If for a given word no translation is found in the bilingual lexicon and no named entities are present in the source and target sentences, the single-word aligner tries to use grammatical tags of source and target words. This is especially the case when the word to align is surrounded with some words already aligned.

Compound-word alignment consists in establishing correspondences between the compound words of the source sentence and the compound words of the target sentences. First, a syntactic analysis is applied on the source and target sentences in order to extract dependency relations between words and to recognize compound words structures. Then, reformulation rules are applied on these structures to establish correspondences between the compound words of the source sentence and the compound words of the target sentence.

In order to use cognates which are present in the source and target sentences, an additional module was added to the single-word aligner. We consider in our approach pairs of words which share the first four characters as cognates. This step uses the transliteration of proper names and detects for example that the proper name "Jackson" and the transliteration of the Arabic word "جاكسون" (Jackson) are cognates. However, this algorithm does not detect pairs of words such as "Blair" and "bleer" (transliteration of the Arabic word "بلير"). To detect these pairs of words, we defined a similarity based on the number of letters in common rather than simply prefixes. This will also detect proper nouns and numerical expressions. The algorithm for cognates detection was adjusted as follows so that it can select only the words of similar size and with a large number of characters in common regardless of the order of these characters. This algorithm uses the following two parameters:

Words_rate = (Number of characters of the short word) / (Number of characters of the long word)

Cognates_rate = (Number of characters in common) / (Number of characters of the short word)

According to this improvement, two words are cognates if *Words_rate* is greater than 0.8 and *Cognates_rate* is greater than 0.5. These two values are fixed empirically.

This algorithm can certainly identify as cognates the word "blair" and the transliteration "bleer "but it also generates errors as is the case of the couple of words "Muhammad" and the transliteration "mahmoud". To reduce the error rate of this module, we added an additional criterion based on the positions of the two words in the source and target sentences.

Table 6 presents results after running all the steps of word alignment process for single and compound words on the French sentence "*M. Blair a imposé des frais d'inscription élevés à l'université qui ont introduit une sélection par l'argent.*" (Mr Blair has imposed high registration fees at the university which introduced a selection by money.) and its Arabic translation " فرض بلير رسوم تسجيل مرتفعة في الجامعة مما ادى الى اختيار الطلاب على قاعدة المال.".

Lemmas of single and compound words of the source language	Lemmas of single and compound words of the target language
Blair (Blair)	بْلِيررَ
imposer (to impose)	فَرَضَ
frais (fees)	رَسْم
inscription (registration)	تَسْجِيل
élevé (high)	مُرْتَفِع
université (university)	جَامِعَة
introduire (introduce)	أدَّى
sélection (selection)	إخْتِيَار
argent (money)	مَال
frais_inscription	رَسْم_تَسْجِيل

Table 6: Result of the alignment of single and compound words.

The word "Blair" was aligned using cognates after transliteration, the words "frais", "élevé" and "introduire" were aligned using grammatical tags and the other single words exist in the bilingual lexicon. The compound word "frais_inscription" was aligned using the reformulation rule *Translation(A.B) = Translation(A).Translation(B)* as follows:

Translation(frais.inscription) = Translation(frais).Translation(inscription) = رَسْم.تَسْجِيل.

5 Experimentation

To evaluate the contribution of the transliteration on the alignment quality of single and compound words, we used two 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 in the training corpus used to extract the translation model of the Moses statistical machine translation system (Koehn et al., 2007).

Because the manual construction of the alignment reference is a difficult and time-consuming task, we conducted a small-scale evaluation based on 283 French-Arabic aligned sentences extracted from the corpus of the ARCADE II campaign. To evaluate the alignment quality, we followed the evaluation framework defined in the shared task on word alignment organized as part of the HLT/NAACL 2003 Workshop on building and using parallel corpora (Mihalcea and Pedersen, 2003). Table 7 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, these results demonstrate that using transliteration improves both precision and recall of word alignment.

Word alignment	Precision	Recall	F-measure
without using transliteration	0.85	0.80	0.82
with the use of transliteration	0.88	0.85	0.86

Table 7: Results of word alignment evaluation.

Certainly, the insufficient size of the corpus used

to evaluate our word aligner does not quantitatively measure the contribution of transliteration but the results clearly indicate an improvement in alignment quality.

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 the 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 10000 pairs of French-Arabic sentences extracted from the ARCADE II corpus. We added to this corpus around 10000 pairs of single and compound words corresponding to the results of our word aligner which integrates transliteration on 500 pairs of French-Arabic sentences. We also specified a language model for the target language using the 10800 Arabic sentences of the ARCADE II corpus.

The performance of the Moses statistical machine translation system is evaluated using the BLEU score on a test corpus composed of 250 pairs of sentences. Note that we consider one reference per sentence. In table 8, we report obtained results.

Training corpora	BLEU
without using transliteration	12.50
with the use of transliteration	12.82

Table 8: Translation results with the BLEU score.

This table shows that the inclusion in the training corpus of word alignment results integrating transliteration reports a gain of 0.32 points BLEU.

It is not obvious at this stage to conclude that this gain in BLEU score induces a significant improvement in translation quality given the low value of this score related to the size of used training corpus (only 10000 pairs of sentences for training the translation model and about 10800 sentences to train the target language model). However, we can easily observe that the transliteration improves the performance of the word aligner whatever the used approach for evaluation: manual or automatic.

6 Conclusion

In this article, we described a transliteration system of proper names from Arabic script to Latin script. This system was used in a word alignment process from a French-Arabic corpus. This process is composed of two steps: First, single words are aligned using an existing bilingual lexicon, named entities, positions and grammatical tags, and second, compound words are aligned using the syntactic dependency relations. This process gives satisfactory and encouraging results when the Arabic transliteration is used to align the names present in the source and target sentences. In future work, we plan, on the one hand, to conduct a large evaluation of our word aligner in order to consolidate the obtained results, and on the other hand, to develop strategies to clean word alignment results in order to construct automatically bilingual lexicons from specialized parallel corpora.

References

Abdulmalik Alsalman, Mansour Alghamdi, Khalid Alhuqayl and Salih Alsubay. 2007. Romanization System for Arabic Names. In *Proceedings of the First International Symposium on Computer and Arabic Language ISCAL – 07*, Riyadh, 214-227.

Bonnie Stalls and Kevin Knight. 1998. Translating Names and Technical Terms in Arabic Text. In *Proceedings of the COLING/ACL Workshop on Computational Approches to Semitic Languages*, Montreal, Québec.

Hasnaa Qunaiir. 2001. Romanizing Arabic names. *Journal er-Riyadh*, 12314.

Joseph Dichy. 2009. La polyglossie de l'arabe illustrée par deux corpus. *M. Bozdemir et L.–J. Calvet (EDS), Politiques linguistiques en méditerranée*, Paris: Honoré Champion, 82-102.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *Proceedings of the 34th ACL Conference*, Madrid, Spain.

Leah S. Larkey, Lisa Ballesteros and Margaret E. Connell. 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland.

Mansour Alghamdi. 2005. Algorithms for Romanizing Arabic names. *Journal of King Saud University* -

Computer and Information Sciences. Riyadh, 17:01-27.

Mathieu Guidère. 2004. *Le traitement de la parole et la détection des dialectes arabes*. Langues stratégiques et défense nationale, Publications du CREC, Saint-Cyr, 53-75.

Melissa. B. Defradas, Rym Hamdi and François Pellegrino. 2004. De la caractérisation linguistique à l'identification automatique des dialectes arabes, In *Proceedings of the MIDL 2004 Workshop*, Paris, France.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris C. 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. In *Proceedings of the 45th ACL Conference, demo session*, Prague, Czech Republic.

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

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

Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat and ChengXiang Zhai. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, Sydney, Australia.

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

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