Mining Named Entity Translation from Non Parallel Corpora

Rahma Sellami*, Fatiha Sadat, Lamia Belguith Hadrich ***

*ANLP Research Group, MIRACL Laboratory,
Sfax University, Sfax, TUNIESIA
rahma.sellami@fsegs.rnu.tn, l.belguith@fsegs.rnu.tn
** UQAM, 201av. President Kennedy,
Montreal, QC, H3X 2Y3, Canada
sadat.fatiha@uqam.ca

Abstract

In this paper, we address the problem of mining named entity translation such as names of persons, organizations, and locations, from non parallel corpora. First, our study concentrates of different forms of named entity translation. Then, we introduce a new framework to extract all named entity translation types from a non parallel corpus. The proposed framework combines surface and linguistic-based approaches. It is language independent and do not rely on any external parallel resources such as bilingual lexicons or parallel corpora. Evaluations show that our approach for mining named entity translations from a non parallel corpus is highly effective and consistently improves the translation quality of Arabic to French machine translation system.

1 Introduction

Machine translation has undergone a major revolution in recent years in the field of Natural Language Processing (NLP). Furthermore, the need for reliable automatic translators is constantly increasing. Indeed, the need to communicate quickly in all languages has become a priority.

Machine Translation (MT) aims to produce a high-quality translation automatically. Though this aim has not yet been achieved, MT research has achieved significant developments.

Named Entities (NEs) are expressions commonly used and appearing frequently in all kinds of texts. These expressions are crucial units for many applications such as information retrieval.

Translating NEs from one language to another opens new perspectives as it can be the basis for new applications, like cross lingual information retrieval, question-answering and machine translation. In machine translation incorrect NE translations not only discard the meaningful information from the original sentences, but also introduce a fuzzy context that degrades the overall translation quality. Moreover, in cross-lingual information retrieval, correct NE translations act as a key query.

Regularly updated documents such as news articles and Web pages usually contain a large number of names. Those names are much more varied than common words, and change continuously. This phenomenon is very problematic for the task of NE translation and plays an important role in boosting the performance and quality of a phrase-based machine translation system. Hence, machine translation systems usually fail to capture those names and then translate them without any specific treatment. For example, many systems translate the Arabic name "داني أبو لوح" (In Buckwalter transliteration: dAny >bw lwH2) into “Danny Abu board”3 or ”Danny Abu plate”4 in English, instead of “Danny Abu lwH” which is the correct transliteration of the Arabic name in our example.

Early researchers focused on mining NE translations from parallel corpora (Kupiec 1993), (Huang et al. 2003). Although such NE translation extraction from parallel corpora can achieve a very high accuracy, exploiting large-scale parallel corpora is not a trivial task. One way to overcome this serious problem is to exploit the much more rich, diverse and readily available non parallel corpora.

Non parallel corpora are widely available on the Web and are regularly updated with new words such as NEs. The

[2] All Arabic transliterations are provided using the Buckwalter transliteration scheme (Buckwalter, 2002)
[3] Translated with the Google Translator on 01/31/2014
[4] Translated with the Reverso Translator on 01/31/2014

most obvious examples of non parallel corpora are the multilingual news produced by news sources such as the BBC, the CNN, Xinhua, the United Nations (UN), Agence France Press (AFP), etc. Also, multilingual encyclopedias, such as Wikipedia (Otero and Lopez 2010; Smith et al. 2010; Sellami et al. 2012a; Sellami et al. 2012b) are considered as comparable or non parallel corpora.

Extracting NE translation pairs from non parallel corpora is a hard task. Even though non parallel corpora contain texts carrying the same information content, those texts are not aligned and exhibit great differences.

Recently, several works have addressed the problem of mining NE transliteration from non parallel corpora (Tao et al. 2006), (Udupa et al. 2008), (Udupa et al. 2009). These methods usually focus only on some types of NE whereas, in practice, most NE translations are not in transliteration. A NE in the source language could be related to a NE in the target one by either transliteration or translation or a combination of both (Bhole et al. 2011).

This paper presents a new study on named entity translation from non parallel corpora and introduces a new framework for extracting NE translation pairs from noisy parallel corpora.

According to Fung et Cheung (2004), a noisy parallel corpus contains the following specifications: most (but not all) of the sentences in the source side have translations in the target side, occurrence frequencies of bilingual word pairs are similar, words have a single translation per corpus and the sentence contexts in two languages of a bilingual word pair are similar.

This framework combines surface and linguistic-based approaches. Unlike the cited previous works, the approach we propose does not rely on any multilingual resource such as bilingual lexicon or a machine translation. Indeed, these resources are scarcely available in some language pairs. As a result, the extracted NE pairs helped to improve the performance of the machine translation systems.

This paper is organized as follows: Section 2 reviews the related research on NE translation detection. In section 3, we present an empirical study of NE translation. A new framework for extracting NE translation pairs from noisy parallel corpora is described in section 4. Section 5 describes the experimental setup. We discuss our evaluation results in section 6. The conclusion and some perspectives are presented in section 7.

2 Related Works

Most previous researches on mining NE translations from non parallel corpora used phonetic similarity to extract NE transliterations. These approaches could not be applied to all types of NE such as organization. For example, the Arabic organization NE "الشركة الأمريكية التطوير والبحث العلمي" (In Buckwalter transliteration: Al$rkp AlOmrykyp lltTwyr wAlbHv AlElmy) should be translated into "The American company for development and scientific research" in English; this translation does not require any transliteration.

Mining NE translation pairs from non parallel corpora was not studied extensively in the literature. (Klementiev et Roth 2006) proposed the use of similarity of temporal distributions for identifying NEs from a comparable corpus. (Shao and Ng 2004) proposed an approach to extract new words, such as NEs and technical terms, and their translations from comparable corpora by combining context and transliteration information.

(Hassan et al. 2007) proposed a language independent approach. First, they align bi-lingual documents, from the comparable corpus, based on their semantic content. The second step extracts two lists of NEs from each pair of aligned documents and then aligns the NEs based on transliteration similarity and phrase based translation similarity.

(Kim et al. 2011) presented the NE translation problem as the matching of two NE graphs extracted from the comparable corpora. A reinforcing method is utilized to reflect relationship similarity and relationship context similarity between NEs.

(Bhole et al. 2011) presented a generative model for extracting equivalents of multi-word NEs from a comparable corpus. (Gupta et al. 2012) proposed a similar approach. The key difference lies in the prior knowledge and the problem formulation. Thus, (Bhole et al. 2011) formulated the problem as a conditional probability of target language multi-word NE alignment for the given source language multi-word NE, while (Gupta et al. 2012) did not assume any prior knowledge of source language multi-word NE and posed the problem as joint probability estimation. You et al. (2013) proposed "selective temporality" as a new feature, as using temporal features might be harmful for translating atemporal entities. They built automatic classifiers to distinguish temporal and atemporal entities then aligned them in separate procedures to boost translation accuracy by 6.1%.

Most previous works require a bilingual lexicon or machine translation system to translate the context of NEs. In this paper, we will present a new framework that does not require any of those resources.

3 Empirical Studies

To understand the various issues in mining NE equivalents from non parallel corpora, we took a random sample of 30 Arabic-French news article pairs from the United Nations corpora for the year 2002. We used the Arabic NE Recognition ArNer (Zribi et al. 2010) to detect the Arabic NEs. ArNer detected 785 NEs of which 187 (24%) were person names, 135 (17%) were location names and 463 (59%)

were organization names. A substantial percentage of the names comprised more than one word: person names 8%, locations 12%, and organizations 92%. For each Arabic NE, we manually identified its equivalent (if any) in the French article. We noticed that each of the studied NE translations usually conformed to one of the following characteristics:

- A single-word Arabic NE can be aligned to a single word French NE.

  Example. جنيف (jnyf) / Genève
- A single-word Arabic NE can be aligned to a multi-word French NE.

  Example. نيويورك (nywywrk) / New York
- A multi-word Arabic NE can be aligned to a multi-word French NE.

  Example. الأمم المتحدة (AlOmm AlmtHdp) / Nations Unies
- A multi-word Arabic NE can be aligned to a single-word French NE.

  Example. الأمم المتحدة (AlOmm AlmtHdp) / ONU

These statistics clearly indicate that multi-word NEs need to be treated as well as single word NE.

4 Extraction of NE Translation Pairs from Non Parallel Corpora

«If an English word e is the translation of a Chinese word c, then the contexts of the two words are similar» (Fung et Yee 1998). Based on this assumption, we introduced a new framework for extracting NE translation pairs from non parallel corpora.

As shown in Figure.1, the proposed approach is composed of three main steps:

1. Aligning slightly the noisy parallel corpus on sentence level.
2. Selecting candidate NE translations based on the sentence alignment.
3. Filtering bad translation pairs based on the context's POS_overlap, type of NEs and distance between NE pair translations.

Step 2 and 3 are executed first to select Arabic to French NE translation pairs and secondly to select French to Arabic NE translation pairs. The intersection of the two sets composes the resulting NE translation pairs.

4.1 Aligning Documents from Noisy Parallel Corpora on a Sentence Level

First, we align the corpora documents based on their meta-information. Then, we align slightly the noisy parallel documents on sentence level. The simple approach of alignment uses a dictionary-based crude translation model. It is based on the Varga algorithm (Varga et al. 2005). First, we align the corpus on sentence level based only on sentence length similarity. Second, a dictionary is extracted from this initial alignment. Finally, we realign the text using a crude word-by-word dictionary-based replacement.

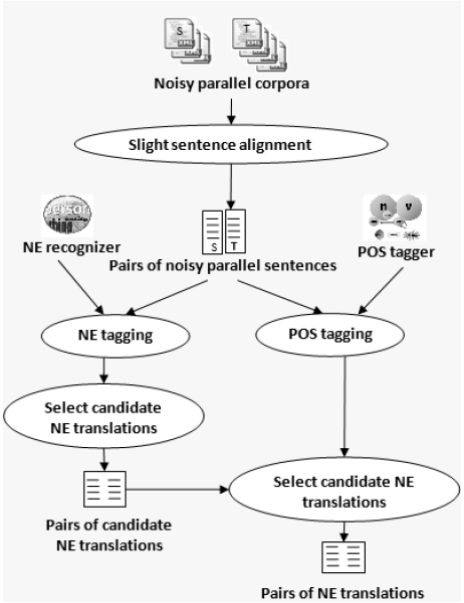

Figure 1. Framework for mining NE translation pairs from noisy parallel corpora.

4.2 Selecting Candidate NE Translation Pairs

We identify the NEs in both source and target sides of our corpora. Since our corpus is slightly aligned on sentence level, searching candidate NE translations while being limited to the sentence alignment can be insufficient. Thus, we search the candidate translations over the corresponding aligned sentence and a window of n sentences.

We calculate a distance d between a NE and its translation candidates based on the sentence alignment. The distance between a NE in sentence Si and a NE in sentence Sj is estimated as d=|j-i|; where i and j are the ranks of these sentences in the aligned noisy parallel corpora.

For each NE in the Arabic language, a list of French translation candidates is constructed and a score d is assigned to each candidate.

The distance d will be subtracted from the score assigned to the translation candidate in order to give priority to the translation pairs belonging to the nearest candidate.

4.3 Filtering Incorrect Translation Pairs

This step is the core of our framework. The previous step selects many translation candidates. To remove bad translations, we propose a filter based on the context's POS_overlap, the type of NEs and the distance between pairs of NEs.

Ideally, the candidate translation list should only contain pairs of NEs having the same type. However, problems arise due to confusion between NE types in the NE extrac-

tion step. To overcome this, we adopt the solution used by (Hassan et al. 2007). A confusion matrix is used to allow NEs of different types. The matrix has an entry for each NE type pair that contains a weighting factor; 1: for the same type, 0: for types that never get confused such as person-location or person-organization, and w: for types that are sometimes confused such as organization-location. The weighting factor will be multiplied by the score assigned to the translation candidate in order to favor translation pairs belonging to the same type.

The context of a NE is constructed from its left and right m POS tags. We compute their POS_overlap: the percentage of POS from the source context that has corresponding POS in the candidate's target NE context. We should mention that we don't take into account the punctuation.

The POS_overlap is the score assigned to the translation candidate. This score will be multiplied by the weighting factor from the confusion Matrix. The candidate pair that has the highest score and that is over than a threshold θ is selected as the best translation.

We calculate the score of a candidate translation as follows:

$$Score = w * POS\_overlap - d * 0{,}1$$

Where w is a weighting factor for the NE confused types. POS_overlap is the percentage of POS from the source context that has corresponding POS in the candidate's target NE context. d is the distance in number of sentences separating a pair of NE candidate.

5 Experimental Setting

In this work, we applied the approach on an Arabic-French noisy parallel corpus to extract NE translation pairs. However, it is worth mentioning that the framework is language independent and could be deployed on any language pairs.

5.1Linguistic Resources

The noisy parallel corpora consist of documents from the ODS of the United Nations (UN). These documents are cleaned from multimedia content, segmented into sentences and converted into XML format (Eisele et Chen., 2010). We used Arabic and French documents of the year 2002. First, we aligned the documents based on their meta-information. A meta-information of a document contains a symbol which is a unique identifier shared between noisy parallel documents. We obtained 4954 pairs of noisy parallel documents, which are considered in our evaluation.

The sentence alignment process was realized with the Hunalign tool (Varga et al. 2005).

We tagged the Arabic and French sides of our noisy parallel corpus with two NE taggers. The Arabic side was tagged with an Arabic NE Recognition system described in (Zribi et al. 2010) while the French side was tagged with a publicly available NER system CasEN5, a Unitex graph cascade for named entity Recognition (Maurel et al 2014).

The Arabic NE Recognizer detects person, location and organization NE types. On the other hand, the French NE Recognizer can detect person, location, organization, time, amount, etc.. So, we concentrated in our work on the person, location and organization type.

In order to compute the POS_overlap, the Arabic side was tagged with the MADA morphological disambiguation system (Habash et al. 2009), and TOKAN, a general Arabic tokenizer (Habash et Sadat 2006). We tokenized the Arabic words with the D3 scheme that splits off clitics as follows: the class of conjunction clitics (w+ and f+), the class of particles (l+, k+, b+ and s+), the definite article (Al+) and all pronominal enclitics. We used the D3 scheme to overcome all problems of agglutination of the Arabic words over the French language. The French side was tagged with the Tree Tagger tool described in (Schmid 1995).

5.2Detection of NE Translations

Based on an empirical study, we fixed the value of n, the window of sentences that contain the translation candidates, on fourteen. So we took seven sentences before and seven sentences after the target sentence aligned to the source sentences containing the NE.

First, for each candidate we attributed a score d which represents the distance between the sentences containing the Arabic NE and the sentences containing the candidate French translation. Second, a score w was attributed by the confusion matrix which indicates if the source type can be matched with the target candidate type. Third, we compute a POS_overlap that represents the percentage of POS tag from the source text that has a corresponding POS tag in the candidate target NE context. We fixed the value of m, the window of POS tag, on ten; five POS tag before the NE and five POS tag after the NE.

These scores (d, w and POS_overlap) were calculated firstly to select the Arabic to French NE translation pairs and secondly to select the French to Arabic NE translation pairs. The intersection of the two sets composes the resulting NE translation pairs.

6 Evaluations

To evaluate the framework performance, we performed experiments on Arabic-French statistical machine translations.

We varied the threshold value θ from 0 to 1and we performed different experimentations. The best result is at-

[5] *http://tln.li.univ-tours.fr/Tln_CasEN.html* visited on August 2013

tained with a threshold value θ equal to 0.8. We don't demonstrate all experiments due to the lack of space.

The baseline system incorporates the "news commentary" bitext as training data. This corpus contains 90753 pairs of sentences. The development data is the test data of nist08. It contains 813 pairs of sentences. The test data set is composed of different issues of the Arabic newspaper "Le Monde Diplomatique"6. It is composed of 423 sentences in the source language (Arabic) with four references. In the test data, 531 Arabic NEs were automatically tagged with the Arabic NE Recognizer (Zribi et al., 2010), among them 30% person, 25% location and 45% organization names.

Word alignment is done with GIZA++ (Och et Ney, 2003). We implemented a 5-gram language model using the SRILM toolkit (Stolcke, 2002). We decode using Moses (Koehn et al., 2007). We tokenized the Arabic side of the training, development and test data using the MADA morphological disambiguation system (Habash et al., 2009), and TOKAN (Habash et Sadat, 2006). French preprocessing of the training, tuning and test data simply included down-casing and separating punctuation from words.

Based on the proposed framework, we extracted 23243 NE translation hypotheses from the Arabic-French UN corpora for the year 2002.

System	Baseline	Baseline+NE	Google Translate
BLEU	32.76	33.05	36.15
OOV	2.37	2.01	0.2

Table1. Improvement of MT quality with NE translation.

We integrated our set of NE translation hypotheses into the baseline system and tested them into the test data. These extracted NE translations from UN corpora were used with the initial "news commentary" parallel corpora. Thus, a training of two language models and two translation models was accomplished using these extracted NE translations and the initial "news commentary" parallel corpora. The resulting system is called Baseline+NE. We report results in terms of case-insensitive 4-gram BLEU scores (Papineni et al., 2002). An appropriate tool for this is the BLEU scoring tool (multi-bleu) which is available in Moses packet. Table 1 shows the improvement of the Arabic-French machine translation quality in terms of BLEU and OOV (Out Of Vocabulary) word, when our NE translation hypotheses were integrated.

When adding NE translations extracted from noisy parallel corpora, an obvious improvement could be observed in the BLEU value as well as in the ratio of OOV words. This can be explained by some Arabic NEs in the test corpora that were messily translated into French in the baseline system. These NEs are correctly translated when we introduce the set of NE translation hypotheses.

Table 1 gives a comparison of Baseline, Baseline+NE and Google Translate7. Google Translate is a translation tool based on statistical models. Its translation memory contains about 200 billion words from the United Nations' documents8. Google Translate gives a better translation than our system. This is due to the big training corpus which is much larger then our training corpus.

Mistranslations of NEs are mainly due to the following causes:

Error in reference files. For example, the person NE "أحمد" (> Hmd) was translated by our system "Ahmed" or he is "Ahmet" in the references. Automatic translation produces a correct correction Arabic word, while the error is in reference files.

Multiple translations assigned to the same place name. For example, the name "تونس" (twns) can be translated as "Tunisia" or "Tunis". These are two correct translations but word source is ambiguous and requires a method of disambiguation based on context. Thus, the word "تونس" (twns) may refer to the country and thus translates into "Tunisia" or the city and thus translates into "Tunis".

7 Conclusion and Perspectives

In this paper, we studied the translation of NEs in non parallel corpora. Then, we proposed a new framework for a language independent NE translation extraction from noisy parallel corpora. Our approach combines surface and linguistic-based approaches and does not rely on any parallel data.

Experiments on Arabic-French language pairs showed that this approach improves substantially the machine translation quality. These results are very encouraging, especially for the Arabic-French pairs of languages. Indeed, these language pairs suffer from the scarcity of parallel data.

In the future, we propose to introduce other features (i.e. features based on transliteration) in order to improve the quality of the extracted NE translation. The framework can be naturally extended to comparable corpora of more than two languages. We intend extracting the NE translation pairs from the entire UN corpora, and the extracted NE will be used to improve sentences and fragment alignment from comparable corpora (i.e., Wikipedia).

[6] *http://www.mondiploar.com/index.php3* visited on August 2013

[7] These experiments were made online dated 01/31/2014

[8] *http://edouard-lopez.com/fac/SciCo%20-%20S5/TAL/projet/TAL%20-%20systran%20vs.%20google%20translate.pdf* visited on August 2013

References

Bhole, A.; Tholpadi, G. and Udupa, R. 2011. "Mining Multi-word Named Entity Equivalents from Comparable Corpora". *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*.

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

Eisele A. and Chen Y. 2010. " MultiUN : A multilingual corpus from United Nation documents". *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10),* Valletta, Malta, May. European Language Resources Association (ELRA).

Fung, P. and Cheung, P. 2004. "Multi-level Bootstrapping for Extracting Parallel Sentences from a Quasi comparable Corpus". *Proceedings of the 20th international conference on Computational Linguistics*.

Fung, P. and Yee, L. Y. 1998. "An IR Approach for Translating New Words from Non parallel, comparable texts". *Proceedings of COLING-ACL*.

Gupta, P. ; Singhal, K. and Rosso, P. 2012. "Multiword Named Entities Extraction from Cross-Language Text Re-use". *CREDISLAS Workshop, LREC2012*. Istanbul, Turkey.

Habash, N; Owen, R. and Ryan, R. 2009. "MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization". Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Habash, N. and Sadat, F. 2006. "Arabic Preprocessing Schemes for Statistical Machine Translation". Prceedings of the 7th Meeting of the North American Chapter of the Association for Computational Linguistics/ Human Language Technologies Conference (HLT-NAACL), New York.

Hassan, A. Fahmy, H. and Hassan, H. 2007. "Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora". *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP), AMML Workshop*.

Huang, F. Vogel, S. and 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*, Vol 15, MultiNER '03, 9–16.

Kim, J. Jiang ; L. Hwang, S. Song, Y. and Zhou, M. 2011. "Mining Entity Translations from Comparable Corpora: A Holistic Graph Mapping Approach". *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM 2011)*. 1295-1304.

Klementiev, A. and Roth, D. 2006. "Slightly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora". *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics. 817–824.

Koehn, P. Hoang; H. Birch, A. Callison-Burch, C. Federico, M. and Bertoldi, N. 2007. "Moses: Open Source Toolkit for Statistical Machine Translation". ACL 2007, demonstration session.

Kupiec, J. , 1993. "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora". *Proceedings of the 31st annual meeting on Association for Computational Linguistics, ACL '93. Association for Computational Linguistics*.

Maurel, D.; Friburger, N; Antoine,J; Eshkol-Taravella, I. and Nouvel, D. 2011. "Cascades de Transducteurs Autour de la Reconnaissance des Entités Nommées". Traitement Automatique des Langues (TAL), Vol 52-1.

Oard, D. W. 1997. "Alternative Approaches for Cross-Language Text Retrieval". *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Articial Intelligence*, 16.

Och F. and Ney H. 2003. "A Systematic Comparison of Various Statistical Alignment Models". Computational Linguistics, 29(1):19–52.

Otero, P. G. and Lopez, I. G. 2010. "Wikipedia as multilingual source of comparable corpora". *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora: from Parallel to Non parallel Corpora*, pages 21{25, Valletta, Malta, May 2010. 16, 45.

Papineni, K.; Roukos, S.; Ward,T. and Zhu,W. 2002. "BLEU: a Method for Automatic Evaluation of Machine Translation". Proceedings of ACL.

Schmid, H. 1995. "Improvements in Part-of-Speech Tagging with an Application to German". Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Sellami, R.; Sadat, F. and Belguith Hadrich, L. 2012a. "Exploiting Wikipedia as a Knowledge Base for the Extraction of Linguistic Resources: Application on Arabic-French Comparable Corpora and Bilingual Lexicons". *Proceedings of the CAASL4 Workshop at AMTA 2012 (Fourth Workshop on Computational Approaches to Arabic Script-based Languages). San Diego, CA*.

Sellami R.; Sadat F. and Belguith Hadrich L. 2012b, "Extraction de lexiques bilingues à partir de Wikipédia*". Atelier de Traitement Automatique des Langues Africaines, JEP-TALN-RECITAL.*

Shao L. and Ng. H. T. 2004. "Mining new word translations from comparable corpora". *In Proceedings of the 20th international conference on Computational Linguistics, COLING '04.* Association for Computational Linguistics, 2004.

Smith, J. R.; Quirk, C.; and Toutanova, K. 2010. "Extracting parallel sentences from comparable corpora using document level alignment". *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403-411, Stroudsburg, PA, USA.

Stolcke A. 2002. "SRILM - an Extensible Language Modeling Toolkit". Proceedings of ICSLP.

Tao Tao; Su youn Yoon; Andrew Fister; Richard Sproat, and Chengxiang Zhai. 2006. "Unsupervised named entity transliteration using temporal and phonetic correlation". *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*.

Udupa Raghavendra, K.; Saravanan, A.; Kumaran, and Jagadeesh Jagarlamudi. 2008, "Mining named entity transliteration equivalents from comparable corpora". *17th ACM conference on Information and knowledge management (CIKM 2008)*, Napa Valley, USA, 1423–1424.

Udupa Raghavendra K. ; Saravanan, A. ; Kumaran, and Jagadeesh Jagarlamudi. 2009. "Mint: A mandhod for effective and scalable mining of named entity transliterations from large comparable corpora". *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009),* Athens, Greece, pages 799–807.

Varga, D; Németh, L. ; Halacsy, P. ; Kornai, A. ; Tron, V. and Nagy, V. 2005. "Parallel Corpora for Medium Density Languages". *Proceedings of RANLP*, Borovets, Bulgaria, 560–596.

You G.; Cha Y.; Kim J.; Hwang S. 2013 "Enriching Entity Translation Discovery using Selective Temporality". *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 201–205, Sofia, Bulgaria, August 4-9 2013.

Zribi, I.; Mezghani Hammami, S; and Belguith Hadrich, L. 2010. "L'apport d'une Approche Hybride pour la Reconnaissance des Entités Nommées en Langue Arabe". *17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.