Arabic-English Text Translation Leveraging Hybrid NER

Emna Hkiri
Latice Laboratory
Faculty of sciences Monastir
Emna.hkiri@gmail.com

Souheyl Mallat
Latice Laboratory
ISIM of Monastir
Souheyl.mallat@gmail.com

Mounir Zrigui
Latice Laboratory
Faculty of sciences Monastir
Mounir.zrigui@fsm.rnu.tn

Abstract

Named Entities (NEs) are a very important part of a sentence and their treatment is a potentially useful preprocessing step for Statistical Machine Translation (SMT). Improper translation of NE lapse the quality of the SMT output and it can hurt sentence's human readability considerably. Dropping NE often causes translation failures beyond the context, affecting both the morphosyntactic formedness of sentences and the word sense disambiguation in the source text. Due to peculiarities of the written Arabic language, the translation task is however rather challenging. In this work, we address the challenging issues of NEs treatment in the context of SMT of Arabic into English texts. We have experimented on three types of named entities which are: Proper names, Organization names and Location names. In this paper, we present integration between machine learning and rule based techniques to tackle Arabic NER problem in attempt to improve the final quality of the SMT system output. We show empirically that each aspect of our approach is important, and that their combination leads to the best results already after integration of NER into SMT. We show improvements in terms of BLEU scores (+4 points) and reduction of out of vocabulary words over a baseline for the News Commentary corpus.

1 Introduction

Named entities recognition is essential for many tasks of natural language processing, whether monolingual or multilingual, as information retrieval or machine translation. In this work, we are interested in the processing of NE in the context of statistical machine translation from Arabic into English, in which the processing of NE poses particular problems. A statistical machine translation (SMT) system learns to translate based on examples of translations already made, extracted from parallel corpus.

In so far as these training corpus are relatively small in size, this raises the question of translation of words that are not seen during training(Hkiri et al., 2015a). This is particularly critical in case of Arabic texts. Arabic is indeed a morphologically complex language (Habash, 2011), many possible forms are rarely observed in the corpus (Heintz, 2008). This requires, at least, implementing morphological analysis to define the source inventory units of the translation system.

A study carried out by Habash (2008) on unknown words in a journalistic corpus for Arabic-English language reports that about 40% of unknown words correspond to proper names. The SMT systems use a default strategy to treat these unknown words; it copies their forms in the output target language. This strategy is sometimes operative especially for person names when the source and target languages use the same alphabet. Unfortunately, this strategy is unsuitable in the case of Arabic into English translation.

To overcome this problem, a common strategy consists to transliterate unknown words in the Latin alphabet (Al-Onaizan and Knight, 2002b), in case of person names and Places (Hermjakob et al., 2008), or even to consult bilingual dictionaries (Hal and Jagarlamudi, 2011).

Treatment of unknown words in Arabic texts in SMT context requires distinguishing between different types of unknown forms, in order to apply differential treatment. In this context, identifying NE appears as a requirement for the text translation. This identification is however difficult in Arabic, in particular because of the lack of distinction between upper and lower case letters which is a valuable indicator to identify proper names in languages using the Latin alphabet. A word form in Arabic texts may refer to different meanings or words according to their context and their diacritics for example the no vowelized word "حافظ" could be a verb (save) for

31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31), pages 124–131
Cebu City, Philippines, November 16-18, 2017

the vowel (FAT'HA) and a personal name for the vowel (Kassra).

Other factors combine to make the identification of NE more challenging. In particular, the use common names as parts of name or surnames or the use of prefixes like (Abd) (servant) associated with a name that describes God. Or, the word Ben (son of) is a part of many names of people from North Africa. The instability of spelling of proper names in various Arabic regions and their diversified transliteration in languages using the Latin alphabet is another source of difficulty. For example "Philippines" is spelled differently: الفليبين or الفيليبين.

In this work, we propose an approach for integrating Named Entities recognition and translation within SMT, which tries to address all these issues at the same time. The objective of applying the treatment of NEs in our statistical machine translation system is to reinforce and improve the quality of Arabic to English text translation. We used DBpedia Linked Data for NER, and the parallel corpus for translation of the recognized NE. For the NER component, we adopted a hybrid approach. We have reproduced the annotator ANNIE incorporated into the GATE tool to serve as baseline rule based component. For machine learning component we exploited the discriminant models using conditional random fields. The rest of the paper is structured as follows. In section2, we will give a literature review of Arabic NER in the SMT. Section 3describes data collection, the architecture of the proposed system and details the main components. Section 4 reports the results of our experiments. In the last section we draw conclusions and discuss some future works.

2 State of the art

For the machine translation (MT) of a text from one natural language to another, named entities require special attention. The MT system should decide whether to translate or transliterate the named entity (Al-Onaizan and Knight, 2002b; Hassan and Sorensen, 2005). In practice, this depends on the type of NE (Chen et al, 2003). For example, personal names tend to be transliterated. The organization names are different, most of them are translated. In contrast, many proper names vary from one language to another. Automatic translation of NEs is one of the most delicate problems; a significant part of mistakes made by the search engines and the most powerful MT tools is caused by NE translation; its bad translation often produce absurd results (Agrawal and Singla. 2012).

Some studies resolve this problem by developing techniques and algorithms for NE transliteration (Santanu, 2010; Hermjakob et al, 2008; Zhang et al, 2011) or by creating domain dictionaries for translation. These last are dictionaries of frequent named entities in a specific area. The quality of the NER system affects the quality of translations (Hkiri et al.,2015). Therefore, the translation of NEs is a fundamental task for most multilingual applications systems (Babych and Hartley, 2003).

There have been few successful attempts on the translation of Arabic named entities. Benajiba (2010) translated directly NE using the automatic alignment of words. Hassan (2007) used the similarity metrics to extract the named entities from bilingual comparable and parallel corpora. Moore (2003) also used the parallel corpus to translate the named entities. The source language in its process is English, so it was based on the initial capitalization to detect proper names. Fehri (2011) translated Arabic entities named using the NooJ platform. Abdul Rauf (2012) improved the translation of entities based on comparable corpus and dictionaries that contain unfamiliar words. Ling (2011) used web links to the translation of NE.

Other published work that uses named entities recognition for machine translation has been directed towards transliterating NEs. The work proposed by Ulf Hermjakob and Kevin Knight et. al. (2008) for Arabic-English translation demonstrates that improvement in translation can be achieved by transliterating NEs instead of trying to translate them. Their work is based on the hypothesis that MT system mistranslates or drops named entities when they do not exist in the training data.

Al-Onaizan (2002a) (2002b) combined translation and transliteration of NE using bilingual and monolingual resources to obtain the best translation of NE. Kashani (2007) transliterates unknown words to improve the performance of the translation system. Jiang (2007) combined the transliteration with data from the web to achieve the best translation of NEs. Azab (2012) reduced the out of vocabulary words of the translation system by automating the translation or transliteration decision from English into Arabic. Abdul Jaleel and Larkey (2003) described their statistical technique of transliteration of the English-Arabic names.

Recently Nasredine and Saadane (2013) developed a system for automatic transliteration of the Arabic proper names in the Latin alphabet.

3 Proposed system

3.1 Data collection

Various linguistic resources are important and necessary in order to develop our Arabic NER system with scope of three different categories of NEs(Hkiri et al.,2016). In the literature, the corpora are commonly used for training, evaluation and comparing with existing systems. The corpora have been cleaned prepared and annotated using our XML format (three named entity tags; one for each NE type person, organization and location).

United nations[1] corpus: is one of the biggest available corpora involving the Arabic language. To obtain our training corpus, we used about 15000 sentences from 2005 Dataset folder. Before using these data files we applied linguistic preprocessing to obtain data in the appropriate automatic processing format.

-ANERcorp[2] dataset is developed by Yassine Benajiba. It is exploited for the training phase of the ML for NER component.

News Commentary 2012[3] Corpus: This corpus consists of political and economic comments from the Project Syndicate website. It has no NE annotations and originally designed to support statistical machine translation in Arabic NLP. Therefore in this research, these datasets have been manually annotated in order to support the NER task. In our study the corpus is used as a reference corpus for NER and SMT evaluation, therefore we extracted and annotated 500 sentences in which we have 350 person names, 410 locations and 151 organizations.

Another type of linguistic resources used is our bilingual NE lexicon[4]: This lexicon is built based on linked datasets of DBpedia and it includes person, place and organization named entities for the couple of Arabic-English languages (Hkiri et al.,2017) (see Table 1).

Named Entities extracted from DBpedia Linked Data	Arabic-English
Person	27480
Organization	17237
Location	4036
Overall	48753

Table1. Bilingual Named Entities lexicon

The described data collection is used for the system development. Our system is based on two relevant components. The NER component is used to detect NE in the source text. This component is based on rule based and machine learning techniques. The second component is dedicated for the Named entities translation (NET component).

3.2 NER component

Both rule based approach and Ml approach have their weakness and strength. By combining them in one hybrid system they may achieve a better performance than operating each of them separately. Our hybrid NER system is a combination of rule-based and ML approaches. The rule-based component is a reproduction of ANIIE system, which is integrated in GATE framework. The ML component uses the CRF model. The system consists of two pipelined components detailed in the following sub-sections:

The rule based component: The rule-based component in our system is a reproduction of the ANNIE system (A Nearly New Information Extraction system) integrated in GATE framework. It is dedicated mainly to the extraction of NE for English. Later the developers have integrated a module for the Arabic language. Nevertheless, the number of Gazetteers for Arabic is much lower compared to that of the English. Time consuming and tedious construction of Arabic Gazetteers lead us to question the way of acquiring an acceptable number of them to ensure better performance of NER system. To overcome this problem we used our bilingual lexicon of NE. In this step, we have exploited the Arabic part of our lexicon; we have mapped our Arabic named entities to predefined gazetteers of GATE as detailed in the following table.

ANNIE/Gate	Predefined entities	Enriched entities from our lexicon
Person	1700	27480
Organization	96	17237

[1] http://www.euromatrixplus.net/multi-un/

[2] http://users.dsic.upv.es/~ybenajiba/downloads.html

[3] http://www.casmacat.eu/corpus/news-commentary.html

[4] https://github.com/Hkiri-emna/Named_Entities_Lexicon_Project

Location	485	4036

Table 2 : Enrichment of predefined Gazetteers of GATE using our lexicon

Moreover, ANNIE is based on the combination of gazetteers and JAPE rules. The idea was to put aside the gazetteers of ANNIE of named entities that we do not need to annotate (such as "URL", "id", "Phone", etc.). We have halved the number of gazetteers. Thus, we simplified the extraction process and we noted a considerable gain in the response time. Similarly, we observed that the JAPE transducer includes a significant n umber of phases (under the .jape file format). Each phase includes a lot of rules, some of which could be inactivated in response to our needs of annotation (annotation rules of the "URL", "id", "Phone", etc.). We believe that these points help to simplify and speed up the base system.

Machine learning based component: The union of rule-based component with the ML component generates the NER hybrid system, which aims to improve the performance of the translation system. The hybridization process is to automatically annotate the test corpus by the rule-based component. The test corpus is annotated again by CRF ++, considering that NE annotated by our rule-based component are correct and CRF ++ is used only to predict areas that have not been annotated. The ML module requires a large amount of annotated data; to do this we used about15000 sentences of United Nations Organization corpus (UN). This corpus is annotated automatically by the rule based module. In addition we used the ANER corpus Benajiba and Rosso.

The latter consists of 4871 sentences. Our supervised ML module uses the Conditional Random Fields model, which is a generalization of Bayesian networks. In our application we used the CRF ++ [5]to annotate sequences of named entities (person, place and organization).

Integrating CRF into Arabic NER : We have used CRF ++ as a development environment for the ML component. This last is based on the set of features, the classification algorithm and the output of the rule-based component. The output of the classification component is used in the prediction phase to generate the final annotation of the NEs. In our study, the output of the hybrid system is analyzed and used to improve the rule-based component.

The selection of features involves selecting a combination of classification functions from the global characteristics space. The features studied in our application are divided into various types: rule-based features, morphological features, POS (morphosyntactic) features and gazetteers features. Each existence x of an element of one of these categories results in testing boolean functions x with each label and each n-gram of the possible labels.

The set of features that are used for NE extraction includes:

Rule-based Characteristics: These contextual elements are the main contribution of the rule-based component to the hybrid system. They come from decisions based on rules defined in terms of a sliding window of size 5 for the immediate right and left neighbors of the candidate word.

Morphological features are derived from the morphological analysis. These characteristics help distinguish the entity named from regular text based on its morphological state. These characteristics are respectively: the aspect, mode and status of the verb, the number of gender, person, voice, whether or not proclitics (such as conjunctions proclitic (Fa), subordinating conjunction (Wa), particles, prepositions (Fi, Bi), the jussive (Li), a marker of future (Sa) negative particles, relative pronouns, etc.

POS feature**:** is the morphosyntactic category of the target word estimated by SAPA tool[6]. This feature allows the classifier to learn the morphosyntactic labels whose named entities occurring with. These labels are: name, number, proper noun, adjective, adverb, pronoun, verb, particle, preposition, conjunctions and punctuation.

Gazetteer features: check the class of the named entity (person, place and organization): a binary function to check if the word (left neighbor / right neighbor of the current word) belongs to predefined Gazetteers categories (person, location, organization). This feature helps to reveal the context of named entities.

Punctuation: This feature indicates whether the word has a point adjacent, for example, at the beginning or the end of the sentence or it is part of an abbreviation. This function allows using the position of text within the classification model

[5] http://crfpp.sourceforge.net/

[6] https://github.com/SouhirG/SAPA

3.3 Named Entities Translation component

The difference of this phase compared to the standard SMT is that we offer hypothesis/proposals of NEs translations to the decoder. During preprocessing step, the Arabic text is segmented and NEs are extracted. Depending on the type of named entities detected, the bilingual lexicon is consulted (Person, location and organizations) in order to avoid ambiguities: a person's name (PERS) can be identical to the name of a street (LOC) as if "الحبيب بورقيبة" this name can be a person's name, airport name or street name. Translations proposals of this named entity, extracted from the bilingual lexicon, are injected in the source text as tags. For example, to name the person "والحبيب بورقيبة" ("Habib Bourguiba"), translations of this NE are proposed to the decoder in the format:
<n translation="Habib Bourguiba || Habib Ben Ali Bourguiba|| Al Habib Bourguiba "> AlHbybbwwrqybp < =n>

4 Experiments and evaluation

4.1 Baseline system

For machine translation, we used our baseline translation system. It integrates GIZA ++ aligner for the training phase. The translation table is formed by aligned segments whose length is up to seven words. The Baseline system was built following the steps in the tutorial of EACL 2009 workshop on statistical machine translation. The difference is that we exploited the UN corpus. The system has been trained and tested on a corpus of about 3.4 million parallels sentences. For Arabic texts, the pipeline of experiences for preprocessing is accomplished on several stages. The first stage is dedicated to the transliteration of texts. The second is devoted to morphological analysis. The next step is normalization. In the last stage, a segmentation of the text is performed to separate the proclitics from the basic word form.
For English text corpus, the main task is tokenization in order to separate punctuation from words. Then, we convert, except for proper names, upper letters by lowercase letters. The final step in this process is data cleaning, it is essential to obtain a high quality translation. In practice, it is difficult to get a perfect set of data or close to perfection. By cleaning our training corpus, we removed:

The source- target repetitive segments, misaligned or identical,

- Too short or too long segments or those who violate the Giza ++ limit ratio,
- Internet links (email, FTP / FTPS, HTTP / HTTPS addresses).

The table below shows the results of preprocessing and cleaning of the UN corpus

	Training Corpus		**test Corpus**	
Language	**Arabic**	**English**	**English**	**Arabic**
N° of tokens	38291993	32645500	8161375	9572998
N° of sentences	1370508	1370508	342627	342627
Avg of tokens /sentence	27,94	23,82	23,82	27,94

Table 3: Statistics: the total number of tokens in Arabic and English corpus

To show the impact of hybridization and the injection of the lexicon as a strengthening NER resource, we conduct an evaluation on News Commentary corpus

4.2 Detection of NE in the News Commentary corpus

This corpus is parallel and it offers us the opportunity to evaluate the translation and recognition of NEs. The News Commentary corpus is extracted from political and economic sites whose topics are close to those of our basic learning corpus, which is extracted from the united nations organization works (UN) .The table below shows performance of baseline NER system, optimized NER system and the hybrid NER system on the News Commentary corpu using standard measures (precison, recall and f-measure)

Named entities		Rule Based NER	Rule Based + NE lexicon	Hybrid NER
Person	P	48.3	80.6	84.3
	R	45.7	79.8	83.34
	F	46.96	80.19	83.81
Organization	P	52.12	71.54	86.24
	R	33.4	59.12	62.5
	F	40.71	64.73	72.47
Location	P	59.5	86.7	89.86
	R	44.6	80.35	89.5
	F	50.98	83.40	89.67
F-measure		46.21	76.10	81.9

Table 4 : Comparison of Baseline, optimized and hybrid NER system on the News Commentary corpus.

-The Baseline system is the rule-based annotator integrated in GATE tool**:** This mode presents modest scores precision, recall and F-measure for all NE classes. This is explained by the lack of Arabic Gazetteers in this annotator. We remind that it is mainly developed for English and later was upgraded for Arabic language processing.

The Baseline system + lexicon is the optimized version: We have enriched the baseline system by our bilingual lexicon. We mapped the Arabic part of the bilingual lexicon to GATE Gazetteers. As a result, we note an improvement in precision for all classes. The strength is the recognition of the place entities, which is attributed to the high coverage of the NE lexicon containing DBpedia datasets.

It is important to note that our system has a good recall for person names, which were more abundant in the UN corpus and in our lexicon (27480 Person NE). Besides, the corpus was a heterogeneous mixture of proper names of persons not only in Arabic countries but also in the continents of Africa, Asia and America ("كوفي أنان" / Kofi Annan, "بان كي مون" / Ban Ki-moon "باراك_أوباما" / Barack Obama). A good percentage of recall for the person NE is encouraging because the named entities of South Asia and America have no phonetic similarity with the names of person in Arabic countries. A detailed review of the results shows that our NER system works poorly for organizations in the corpus, in fact, our system does not effectively manages acronyms or abbreviations.

The Hybrid system is the final version. The results show an improvement in overall F-measure of NE classes (+5 points) compared to previous results. Note that the hybrid model improves recognition of all NEs and especially the recognition of places since the lexicon-based system has better performance on the recognition of places.

4.3 Evaluation of the impact of NER on the SMT system

For SMT, we used the Moses decoder that integrates GIZA ++ aligner used in the training phase. The translation table generated consists of segments up to seven words. The SAPA tool is used for pretreatment of Arabic texts.

We remind that the basic principle of our translation method is to propose translations of NE to the decoder. During the preprocessing phase, the Arabic text is pretreated and the named entities are annotated. Depending on the class or category of NE detected, the bilingual lexicon is consulted (people, places and organizations) to avoid ambiguities in polysemic entities. Proposals of translations extracted from the NE lexicon are injected in the text to be translated in two modes of translation (inclusive and exclusive). The annotation of NE in source and target News Commentary corpus allows us to automatically evaluate the quality of the translation of NEs. We evaluated three modes of translation summarized below.

The Default mode: As the name indicates, in this mode no treatment of named entities is accomplished. It presents the translation generated by our baseline system.

The Exclusive mode: In this mode, only proposals of translations offered by the lexicon are considered in the calculation of the best translation score.

The Inclusive mode: In this mode the translations provided by the lexicon and the translations from the translation table are considered in calculating the score.

We remind that we have achieved learning on the UN corpus and for evaluation experiments we used the News Commentary corpus. The results of the evaluation are in terms of BLEU score. Table below shows the translation results in the three modes of translation. The total of out of vocabulary words (OOV) is also presented.

	BLEU	Mots OOV
Default	32.35	145
Inclusive	36.2	115
Exclusive	32.14	115

Table 5: BLUE and OOV scores for the Arabic-English translation of 500 sentences of the News Commentary2012 corpus.

Comparing the exclusive mode by default mode, we notice a slight decrease in BLEU score. This is because some translations proposed by our lexicon differ from those of the reference. A more detailed analysis shows that our lexicon does not provide incorrect translations, but they are different from those of the reference. An example is the translation "منظمة حلف شمال الأطلسي" in our lexicon the word is translated by North Atlantic Treaty Organization, while for the reference is abbreviated to "NATO". In some cases, our translations correct those of the reference as an example of the place "سيريلانكا" it is translated in the reference by Srilanka, while our system, it says Sri Lanka. Also, using this mode some named entities translations are improved. They are translated correctly by our system but they are incorrect for the Baseline system (default). Cite the example of NE " الرئيس جورج بوش الأب " it is translated by the Baseline system "President Bush" whereas the reference is "president George Herbert Walker Bush"

The exclusive mode does not improve translation quality, but it affects the rate of OOV words. The percentage declined, he passed by 145 in the default mode to reach 115 for the exclusive mode.

According to the BLEU score, inclusive mode is the best, with a decrease of OOV words. Therefore, we can say that the idea of integrating translations extracted from the bilingual lexicon, improves translation quality while ensuring, as shown above, better coverage of the named entities.

For evaluating the translation of named entities, we will limit to the inclusive mode. The rate of correctly translated NEs was calculated for each class on the test corpus. The calculation is made by comparing NE translated to those in the reference.

	Person	Location	Organization
Default	53.36%	73.50%	46.42%
Inclusive	80.05%	86.31%	62.80%

Table 6: evaluation of effect of NE Translation on the News Commentary corpus.

Using the inclusive method improves the rate of NE translated correctly compared to the default system. These quantitative results show that the use of the lexicon does affect the translation of named entities, although this is not always reflected by a significant increase in BLEU score.

5 Conclusion

In this work, we addressed the main problems of NE integration into an SMT system. Our approach integrates a hybrid NER system, and allows choosing adapted NE translations for each NE. In conclusion, it can be said that using NEs does help in providing better SMT. we did improve the BLEU score over baseline system, a number of translated sentences show improvement with the use of these techniques. There was a considerable reduction in mistranslation and dropping of NEs. This helped enhancing human readability as well. Analysis of our models also revealed a number of insights and scopes for further improvement. There is also a space for using different ML techniques other than CRF, and how this will impact on the performance of the NER system

References

AbdulJaleel Nasreen, and Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross language information retrieval. Proceedings of the twelfth international conference on Information and knowledge management. ACM,

Abdul Rauf Sadaf. 2012. Efficient corpus selection for Statistical Machine Translation. PhDthesis. Le Mans, France.

Al-Onaizan Yaser, Knight Kevin. 2002a. Machine transliteration of names in Arabic text. Proc. of the ACL-02 workshop on Computational approaches to semitic languages: 1-13.

Al-Onaizan Yeser, Knight Keven 2002b. Translating named entities using monolingual and bilingual resources. Proc. of the 40th Annual Meeting on ACL, ACL '02, Philadelphia, PA, USA: 400-408.

Agrawal Neeraj and Singla Ankush. 2012. Using named entity recognition to improve machine

translation. Technical report, Standford University, NaturalLanguage Processing..

Babych, Bogdan, and Anthony Hartley.2003. Improving machine translation quality with automatic named entity recognition. In: Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. EAMT '03. Budapest, Hungary: Association for Computational Linguistics: 1-8.

Benajiba Yassine, Zitouni Imed. 2010. Enhancing mention detection using projection via aligned corpora. In: Proceedings of the 2010 conference on empirical methods in natural language processing, Cambridge. Association for Computational Linguistics:993-1001.

Chen, Hsin-Hsi, Changhua Yang, and Ying Lin. 2003.Learning formulation and transformation rules for multilingual named entities. Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition- 15(1). Association for Computational Linguistics.

Fehri Hela., Haddar Kais., Ben Hamadou Abdelmajid. 2011. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model.2011. In M. Constant, A. Maletti, A. Savary (eds), FSMNLP, 9th International Workshop:134-142,.

Habash Nizar.2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies:57-60.

Habash Nizar, 2011.Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies, organ & Claypool Publishers.

Daumé III, Hal, and Jagadeesh Jagarlamudi. 2011 . Domain adaptation for machine translation by mining unseen words, Proceedings of the 49th Annual Meeting of the ACL :HLT : short papers - Volume 2, HLT '11, ACL, Stroudsburg, PA, USA, p. 407-412.

Hammersley John M. and Peter Cliford. 1971. Markov fields on finite graphs and lattices.

Hassan, Ahmed, Haytham Fahmy, and Hany Hassan. 2007. Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora. In: Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP '07).

Hassan, Hany, and Jeffrey Sorensen. 2005. An integrated approach for Arabic-English named entity translation. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Semitic '05, ACL:87-93,

Heintz Ilana..2008. Arabic Language Modeling with Finite State Transducers. In Proc. of the ACL-08: HLT Student Research Workshop, ACL, Columbus, Ohio, p. 37-42.

Hermjakob, Ulf, Kevin Knight, and Hal Daumé III. 2008. Name Translation in Statistical Machine Translation - Learning When to Transliterate. Proc. of ACL-08 : 389-397

Hkiri Emna, Mallat Souheyl, and Zrigui Mounir.2017.Constructing a Lexicon of Arabic-English Named Entity using SMT and Semantic Linked Data. IAJIT, vol.6, to appear November 2017.

Hkiri Emna, Mallat Souheyl, and Zrigui Mounir, 2016. Events Automatic Extraction from Arabic Texts. IJIRR, vol. 6(1), pp. 36-51.

Hkiri Emna, Mallat Souheyl, and Zrigui Mounir.2015.Improving coverage of rule based NER systems," ICTA, pp.1-6.

Hkiri Emna, Mallat Souheyl, and Zrigui Mounir.2015a.Automating Event Recognition for SMT Systems. ICCCI, pp.494-502.

Jiang, Long, Zhou, Ming, Chien, Lee-Feng. 2007. Named Entity Translation with Web Mining and Transliteration. In: Proceedings of the 20th International Joint Conferenceon Artificial Intelligence:1629-1634

Mehdi M. Kashani, Simon Fraser, Eric joanis, George Foster, Fred Popowich 2007. Integration of an Arabic transliteration module into a statistical machine translation system.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Ling, Wang, Calado, Pável, Martins, Bruno 2011. Named Entity Translation using Anchor Texts. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT).

Moore Robert C. 2003. Learning Translations of Named-Entity Phrases from Parallel Corpora. In: In Proc. Of Eacl: 259-266.

Nasredine Semmar, Saadane Houda.2013. Using Transliteration of Proper Names from Arabic to Latin Script to Improve English-Arabic Word Alignment. IJCNLP :1022-1026, 2013.

PAL, Santanu, Kumar Naskar, Sudip, Pecina, Pavel, 2010. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation. In: Proceedings of the COLING 2010 Workshop on Multiword Expressions: from Theory to Applications (MWE 2010):46-54,

Zhang, Min, Haizhou Li, Kumaran Ming L. 2011. Report of NEWS2011 Machine Transliteration Shared Task. In: Proceedings of 2011 Named Entities Workshop. Chang Mai, Thailand.