Dudley North visits North London: Learning When to Transliterate to Arabic

Mahmoud Azab **Houda Bouamor** **Behrang Mohit** **Kemal Oflazer**
Carnegie Mellon University
P.O. Box 24866, Doha, Qatar
{mazab, hbouamor, behrang, ko}@qatar.cmu.edu

Abstract

We report the results of our work on automating the transliteration decision of named entities for English to Arabic machine translation. We construct a classification-based framework to automate this decision, evaluate our classifier both in the limited news and the diverse Wikipedia domains, and achieve promising accuracy. Moreover, we demonstrate a reduction of translation error and an improvement in the performance of an English-to-Arabic machine translation system.

1 Introduction

Translation of named entities (NEs) is important for NLP applications such as Machine Translation (MT) and Cross-lingual Information Retrieval. For MT, NEs are major subset of the out-of-vocabulary terms (OOVs). Due to their diversity, they cannot always be found in parallel corpora, dictionaries or gazetteers. Thus, state-of-the-art of MT needs to handle NEs in specific ways. For instance, in the English-Arabic automatic translation example given in Figure 1, the noun "North" has been erroneously translated to "الشمالية /Al$mAlyp " (indicating the north direction in English) instead of being transliterated to "نورث / nwrv".

As shown in Figure 1, direct translation of invocabulary terms could degrade translation quality. Also blind transliteration of OOVs does not necessarily contribute to translation adequacy and may actually create noisy contexts for the language model and the decoder.

English Input:	Dudley **North** was an English merchant.
SMT output:	كان دودلي **الشمالية** تاجر الإنجليزية. kAn dwdly **Al$mAlyp** tAjr AlInjlyzyp.
Correct Translation:	كان دودلي **نورث** تاجر إنجليزي. kAn dwdly **nwrv** tAjr Injlyzy.

Figure 1: Example of a NE translation error.

An intelligent decision between translation and transliteration should use semantic and contextual information such as the type of the named-entity and the surrounding terms. In this paper, we construct and evaluate a classification-based framework to automate the translation vs. transliteration decision. We evaluate our classifier both in the limited news and diverse Wikipedia domains, and achieve promising accuracy. Moreover, we conduct an extrinsic evaluation of the classifier within an English to Arabic MT system. In an in-domain (news) MT task, the classifier contributes to a modest (yet significant) improvement in MT quality. Moreover, for a Wikipedia translation task, we demonstrate that our classifier can reduce the erroneous translation of 60.5% of the named entities.

In summary our contributions are: (a) We automatically construct a bilingual lexicon of NEs paired with the transliteration/translation decisions in two domains.[1] (b) We build a binary classifier for transliteration and translation decision with a promising accuracy (c) We demonstrate its utility

[1]The dataset can be found at http://www.qatar.cmu.edu/~behrang/NETLexicon .

Proceedings of NAACL-HLT 2013, pages 439–444,
Atlanta, Georgia, 9–14 June 2013.

within an MT framework.

2 Learning when to transliterate

We model the decision as a binary classification at the token level. A token (within a named-entity) gets translation or transliteration label. In "*Dudley North*" and "*North London*", our classifier is expected to choose transliteration of "North" in the former case, as opposed to translation in the latter. The binary decision needs to use a rich set of local and contextual features. We use the Support Vector Machines as a robust framework for binary classification using a set of interdependent features.[2] We build two classifiers: (a) **Classifier** $\mathbf{C}_{news}$, trained on a large set of distinct NEs extracted from news-related parallel corpora; and (b) **Classifier** $\mathbf{C}_{diverse}$, trained on a combination of the news related NEs and a smaller set of diverse-topic NEs extracted from Wikipedia titles. We evaluate the two classifiers in both news and the diverse domains to observe the effects of noise and domain change.

2.1 Preparing the labeled data

Our classifier requires a set of NEs with token-level gold labels. We compile such data from two resources: We heuristically extract and label parallel NEs from a large word aligned parallel corpus and we use a lexicon of bilingual NEs collected from Arabic and Wikipedia titles. Starting with a word aligned parallel corpus, we use the UIUC NE tagger (Ratinov and Roth, 2009) to tag the English sentences with four classes of NEs: Person (PER), Location (LOC), Organization (ORG) and Miscellaneous (MISC). Furthermore, we use the word alignments to project and collect the span of the associated Arabic named-entities. To reduce the noisy nature of word alignments, we designed a procedure to clean up the noisy Arabic NE spans by POS verification, and heuristically filtering impossible items (e.g. verbs). This results in a bilingual lexicon of about 57K named-entity pairs. The distribution of NEs categories is reported in Table 1.

To train and evaluate the $\mathbf{C}_{diverse}$ classifier, we expand our labeled data with Wikipedia NEs using the cross-lingual hyperlinks. Wikipedia article titles often correspond to NEs (Kazama and Torisawa, 2007) and have been already used in different works for NEs recognition (Nothman et al., 2013) and disambiguation (Cucerzan, 2007). We improve the Arabic-English Wikipedia title lexicon of Mohit et al. (2012) and build a Wikipedia exclusive lexicon with 4K bilingual entities. In order to test the domain effects, our lexicon includes only NEs which are not present in the parallel corpus. The statistics given in Table 1 demonstrate different nature of the labeled datasets. The two datasets were labeled semi-automatically using the transliteration similarity measure (Fr_{score}) proposed by Freeman et al. (2006), a variant of edit distance measuring the similarity between an English word and its Arabic transliteration. In our experiments, English tokens having an $\text{Fr}_{score} > 0.6$ are considered as transliteration, others having $\text{Fr}_{score} < 0.5$ as translation. These thresholds were determined after tuning with a held out development set. For tokens having Fr_{score} between 0.5 and 0.6, the decision is not obvious. To label these instances (around 5K unique tokens), we manually transliterate them using Microsoft Maren tool.[3] We again compute the Fr_{score} between the obtained transliteration, in its Buckwalter form and the corresponding English token and use the same threshold to distinguish between the two classes. Some examples of NEs and their appropriate classes are presented in Table 2.

	PER	LOC	ORG	MISC
News$_{/57K}$	43.0%	10.0%	40.0%	7.0%
Wiki$_{/4K}$	73.0%	19.0%	2.5%	5.5%

Table 1: Distribution of the four NE categories used in 57K News and 4K Wiki datasets.

Transliteration	Translation
Minnesota ↔ مينيسوتا/mynyswta : 0.77	Agency ↔ وكالة/wkAlp : 0.33
Fluke ↔ فلوك/flwk : 0.57	Islamic ↔ الاسلامية/AlAslAmyp : 0.55

Table 2: Examples of NEs labeled using Freeman Score.

2.2 Classification Features

We use a total of 32 features selected from the following classes:

Token-based features: These consist of several features based on the token string and indicate

[2] We use the LIBSVM package (Chang and Lin, 2011).

[3] http://afkar.microsoft.com/en/maren

whether the token is capital initial, composed entirely of capital letters, ends with a period (such as Mr.), contains a digit or a Latin number (e.g. Muhammad II) or contains punctuation marks. The string of the token is also added as a feature. We also add the POS tag, which could be a good indicator for proper nouns that should mainly be transliterated. We also check if the token is a regular noun in the WORDNET (Fellbaum, 1998) which increases its chance of being translated as opposed to transliterated.

Semantic features: These features mainly indicate the NE category obtained using an NE tagger. We also define a number of markers of person (such as Doctor, Engineer, etc.) and organization (such as Corp.) names. We used the list of markers available at: `http://drupal.org/node/1439292`, that we extended manually.

Contextual features: These features are related to the token's local *context within the NE*. These include information about the current token's surrounding tokens, its relative position in the NE (beginning, middle or end). Another feature represents the length of the NE in number of tokens.

2.3 Experiments

We train two classifiers and tune their parameters using a held out development set of 500 NEs drawn randomly from the news parallel corpus. We use 55k NEs from the same corpus to train the $\mathbf{C}_{news}$ classifier. Furthermore, we train the $\mathbf{C}_{diverse}$ classifier cumulatively with the 55K news NEs and another 4600 NEs from Wikipedia titles.

The classifiers are evaluated on three different datasets: $\mathbf{Test}_{News}$ which consists of 2K of NEs selected randomly from the news corpus, $\mathbf{Test}_{Wiki}$ consisting of 1K NEs extracted from the Wikipedia and $\mathbf{Test}_{Combination}$, an aggregation of the two previous sets. We manually reviewed the labels of these test sets and fixed any incorrect labels. Table 3 compares the accuracy of the two classifiers under different training and test data settings. Starting with a majority class baseline, our classifiers achieve a promising performance in most settings. The majority class for both classifiers is the *translation* which performs as a baseline approach with an accuracy equal to the distribution of the two classes. We also observe that the addition of a small diverse training set in $\mathbf{C}_{diverse}$ provides a relatively large improvement (about 2%) when tested on Wikipedia. Finally, Figure 2 illustrates the contribution of different classes of features on our diverse classifier (evaluated on $\mathbf{Test}_{Wiki}$). We observe a fairly linear relationship between the size of the training data and the accuracy. Furthermore, *we observe that the features describing the category of the NE are more important than the token's local context*. For example, in the case of "Dudley North" and "North London", the most effective feature for the decision is the category of the named entities.

	$\mathbf{Test}_{News}$	$\mathbf{Test}_{Wiki}$	$\mathbf{Test}_{Combination}$
Baseline	56.70	57.09	56.89
$\mathbf{C}_{news}$	90.40	84.10	88.64
$\mathbf{C}_{diverse}$	**90.42**	86.00	89.18

Table 3: Accuracy results for the two classifiers and the baseline on the three test datasets

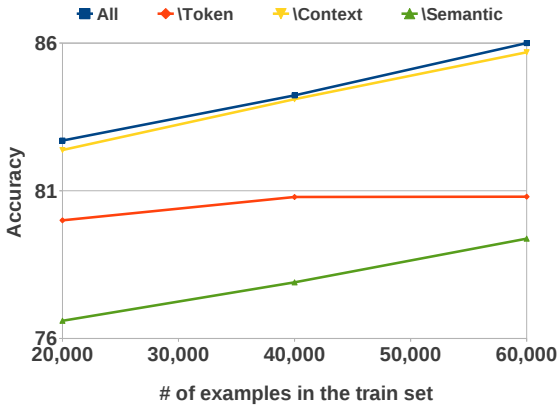

Figure 2: Learning curves obtained on Wiki dataset by removing features individually.

3 Extrinsic MT evaluation

We evaluate the effects of the classifier on an English to Arabic statistical MT system. Our first evaluation focuses on the utility of our classifier in preventing erroneous translation of NEs which need to be transliterated. In the following experiments we use $\mathbf{C}_{news}$ classifier. In order to experiment with a diverse set of NEs, we conducted a study on a small corpus (98,197 terms) of Wikipedia articles from a

diverse set of topics. We use 10 Wikipedia articles describing: Anarchism, Artemis, Buddhism, Isfahan, Shawn Michaels, Turkey, etc. We first use our classifier to locate the subset of NEs which should be transliterated. An annotator validates the decision and examines the phrase table on the default MT decision on those NEs. We observe that out of 1031 NE tokens, 624 tokens (60.5%) which would have been translated incorrectly, are directed to the transliteration module.

Finally, we deploy the transliteration classifier as a pre-translation component to the MT system.[4] Our MT test set is the MEDAR corpus (Maegaard et al., 2010). The MEDAR corpus consists of about 10,000 words English texts on news related to the climate change with four Arabic reference translations. Due to the lack of non-news English-Arabic corpus, we have to limit this experiment only to the news domain. However, we expect that many of the NEs may already exist in the training corpus and the effects of the classifier is more limited than using a diverse domain like Wikipedia. We automatically locate the NEs in the source language sentences and use the classifier to find those which should be transliterated. For such terms, we offer the transliterated form as an option to the decoder aiming to improve the decoding process. For that a human annotator selected the transliterations from the suggested list that is provided by the automatic transliterator (Maren) without any knowledge of the reference transliterations.

Table 4 shows the impact of adding the classifier to the SMT pipeline with a modest improvement. Moreover, a bilingual annotator examined the automatically tagged NEs in the MT test set and labeled them with the translation vs. transliteration decisions. Having such gold standard decisions, we evaluated the classifier against the MT test set. The classifier's accuracy was 89% which is as strong as the earlier intrinsic evaluation. The false positives are 5% which represents around 12.6% of the total errors.

	MT Baseline	**MT Baseline + Classifier**
BLEU	16.63	**16.91**

Table 4: Results of the extrinsic usage of the classifier in SMT

The following example shows how our classifier prevents the MT to choose a wrong decoding for the NE *Python* (being transliterated rather than translated). Moreover, the MT system transliterates the term *Monty* that is unknown to the underlying system. Such entities tend to be unseen in the standard news corpora and consequently unknown (UNK) to the MT systems. Using our classifier in such conditions is expected to reduce the domain gap and improve the translation quality.

English Input: The British comedy troupe **Monty Python**.

Baseline MT: الفرقة الكوميدية البريطانية **UNK** افعي.
Alfrqp Alkwmydyp AlbryTAnyp **UNK AfEY**

MT+Classifier: الفرقة الكوميدية البريطانية **مونتي بايثون** .
Alfrqp Alkwmydyp AlbryTAnyp **mwnty bAyvwn**.

4 Related work

A number of efforts have been made to undertake the NE translation problem for different language pairs. Among them some use sequence of phonetic-based probabilistic models to convert names written in Arabic into the English script (Glover-Stalls and Knight, 1998) for transliteration of names and technical terms that occurs in Arabic texts and originate in English. Others rely on spelling-based model that directly maps an English letter sequence into an Arabic one (Al-Onaizan and Knight, 2002a). In a related work, Al-Onaizan and Knight (2002b) describe a combination of a phonetic-based model and a spelling-based one to build a transliteration model to generate Arabic to English name translations. In the same direction, Hassan et al. (2007) extracted NE translation pairs from both comparable and parallel corpora and evaluate their quality in a NE translation system. More recently, Ling et al. (2011) propose a Web-based method that translates Chinese NEs into English. Our work is similar in its general objectives and framework to the work pre-

[4]The baseline MT system is the MOSES phrase-based decoder (Koehn et al., 2007) trained on a standard English-Arabic parallel corpus. The 18 million parallel corpus consists of the non-UN parts of the NIST corpus distributed by the Linguistic Data Consortium. We perform the standard preprocessing and tokenization on the English side. We also use MADA+TOKAN (Habash et al., 2009) to preprocess and tokenize the Arabic side of the corpus. We use the standard setting of GIZA++ and the grow-diagonal-final heuristic of MOSES to get the word alignments. We use a set of 500 sentences to tune the decoder parameters using the MERT (Och, 2003). We use El Kholy and Habash (2010) detokenization framework for the Arabic decoding. We evaluate the MT system with the BLEU metric (Papineni et al., 2002).

sented by Hermjakob et al. (2008), which describes an approach for identifying NEs that should be transliterated from Arabic into English during translation. Their method seeks to find a corresponding English word for each Arabic word in a parallel corpus, and tag the Arabic words as either NEs or non-NEs based on a matching algorithm. In contrast, we tackle this problem in the reverse direction (translating/transliterating English NEs into Arabic). We also present a novel binary classifier for identifying NEs that should be translated and those that should be transliterated.

5 Conclusion and future work

We reported our recent progress on building a classifier which decides if an MT system should translate or transliterate a given named entity. The classifier shows a promising performance in both intrinsic and extrinsic evaluations. We believe that our framework can be expanded to new languages if the required data resources and tools (mainly parallel corpus, Named Entity tagger and transliteration engine) are available. We plan to expand the features and apply the classifier to new languages and conduct MT experiments in domains other than news.

6 Acknowledgements

We thank Nizar Habash and colleagues for the MADA, Arabic detokenization and the transliteration similarity software and also their valuable suggestions. We thank anonymous reviewers for their valuable comments and suggestions. This publication was made possible by grants YSREP-1-018-1-004 and NPRP-09-1140-1-177 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Yaser Al-Onaizan and Kevin Knight. 2002a. Named-Entity translation. In *Proceedings of HLT*, San Francisco, USA.

Yaser Al-Onaizan and Kevin Knight. 2002b. Translating Named Entities Using Monolingual and Bilingual Resources. In *Proceedings of ACL*, Philadelphia, USA.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Silviu Cucerzan. 2007. Large-Scale Named-Entity Disambiguation Based on Wikipedia Data. In *Proceedings of EMNLP-CoNLL*, Prague, Czech Republic.

Ahmed El Kholy and Nizar Habash. 2010. Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of LREC*, Valletta, Malta.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. *The MIT Press*.

Andrew Freeman, Sherri Condon, and Christopher Ackerman. 2006. Cross Linguistic Name Matching in English and Arabic. In *Proceedings of NAACL*, New York City, USA.

Bonnie Glover-Stalls and Kevin Knight. 1998. Translating Named and Technical Terms in Arabic Text. In *Proceeding of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, Montreal, Canada.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+Tokan: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora. In *Proceedings of RANLP*, Borovets, Bulgaria.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name Translation in Statistical Machine Translation - Learning When to Transliterate. In *Proceedings of ACL-HLT*, Columbus, Ohio.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named-Entity Recognition. In *Proceedings of EMNLP-CoNLL*, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola 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 ACL: Demo session*, Prague, Czech Republic.

Wang Ling, Pavel Calado, Bruno Martins, Isabel Trancoso, and Alan Black. 2011. Named-Entity Translation using Anchor Texts. In *Proceedings of IWSLT*, San Francisco, USA.

Bente Maegaard, Mohamed Attia, Khalid Choukri, Olivier Hamon, Steven Krauwer, and Mustafa Yaseen. 2010. Cooperation for Arabic Language Resources and Tools–The MEDAR Project. In *Proceedings of LREC*, Valetta, Malta.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-

Oriented Learning of Named Entities in Arabic Wikipedia. In *Proceedings of EACL*, Avignon, France.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning Multilingual Named Entity Recognition from Wikipedia. *Artificial Intelligence*, 194(0):151 – 175.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, USA.

Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of CONLL*, Boulder, USA.