Unsupervised False Friend Disambiguation Using Contextual Word Clusters and Parallel Word Alignments

Maryam Aminian, Mahmoud Ghoneim, Mona Diab
Department of Computer Science
The George Washington University
Washington, DC
{aminian,mghoneim,mtdiab}@gwu.edu

Abstract

Lexical false friends (FF) are the phenomena where words that look the same, do not have the same meaning or lexical usage. FF impose several challenges to statistical machine translation. We present a methodology which exploits word context modeling as well as information provided by word alignments for identifying false friends and choosing the right sense for them in the context. We show that our approach enhances SMT lexical choice for false friends across language variants. We demonstrate that our approach reduces word error rate (WER) and position independent error rate (PER) for Egyptian-English SMT by 0.6% and 0.1% compared to the baseline.

1 Introduction

False friends (FF), aka faux amis, are words in two or more language variants that are orthographically and/or phonetically similar but do not convey the same meaning (Brown and Allan, 2010). FF sense divergence is one of the main sources of performance degradation in statistical machine translation (SMT) systems. These words are frequently observed when the underlying distribution of the test set is different from that of the train data. In other words, the sense of a particular word in the input sentence varies from all observed senses of that word in the train data. Thus, SMT may choose the target language translation which is considered inappropriate based on the context.

Standard form of a language has different informal spoken varieties which are known as dialects. For instance, standard form of Arabic has different dialects (Habash, 2010). These dialects typically share a set of cognates that could bear the same meaning in both varieties or only be shared homographs but serve as false friend. The usage of dialects in textual social media and communication channels is rapidly increasing. On the other hand, usually there is not enough dialectal parallel data to train the translation model and build stand alone machine translation systems for dialects. However, the standard official forms of language usually have a wealth of resources and tools that can be adapted to dialects of that language.

The main goal of this paper is to enhance dialectal SMT performance without any in-domain training data. We move towards this goal by performing a pre-processing phase which includes, 1) identifying false friends in the input sentence and, 2) replacing them with an appropriate equivalent from standard language which bears the same meaning. By doing this, we benefit from availability of standard parallel data to choose a more accurate target translation for the false friends.

We aim to identify false friends without any labeled training data. We then try to choose equivalents from the standard language for the identified false friends. We exploit a classifier for identifying false friends and designing a word sense disambiguator for finding the best equivalent from the standard language. We employ unsupervised word alignment from parallel text and a taxonomy-based semantic similarity measure (Wu and Palmer, 1994) to automatically acquire training data for the FF identifier. Our word sense disambiguator benefits from unsupervised word clusters to model the context. We obtain word clusters from a large monolingual text in the standard language. Training the model only involves counting the coocurrences of

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 39–48,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

each word with word clusters for different context definitions. During decoding (disambiguation), for a word in a sentence, we estimate the likelihood for each equivalent of that word given word clusters in its surrounding context.

We evaluate our method on Egyptian (EGY) to English (EN) SMT using a translation model trained on Modern Standard Arabic (MSA). We show that our approach improves EGY-to-EN SMT lexical choice and reachs 0.6% and 0.1% reduction in word error rate (WER) and position-independent error (PER) (Tillmann et al., 1997) over the baseline respectively. In summary, the main contributions of this paper are: 1) designing a FF identifier with a supervised classifier trained on automatically acquired labeled data, 2) designing a disambiguator for replacing FF with their equivalent standard form and 3) improving the SMT lexical choice on dialectal data without using any in-domain parallel data to train SMT model.

The remainder of this paper is organized as follows: We give a literature overview in §2. We then detail our approach in §3 . We present t experiments in §4 and discuss the results in §5. We finally make conclusions in §6.

2 Related Work

There have been several studies for identifying false friends which benefit from parallel data to measure semantic similarity of words (Frunza and Inkpen, 2006; Nakov et al., 2009; Inkpen et al., 2005; Kondrak, 2001; Mitkov et al., 2007). Some other studies such as (Nakov et al., 2007; Schulz et al., 2004; Nakov et al., 2009; Mulloni et al., 2007) exploit distributional semantics to identify false friends. These methods hypothesize that words occurring in similar contexts tend to be semantically similar. Methods leveraging this idea usually use vector space models to show the local context of the target word. Context can be modeled either with a window of a certain size around the target word e.g. (Nakov et al., 2009) and (Schulz et al., 2004) or words in a particular syntactic relationships with the target word e.g. (Mulloni et al., 2007).

The most comparable work to our false friend identification approach is the work done by Mitkov et al. (2007) which uses both distributional semantic evidences extracted from monolingual data and bilingual hints obtained from comparable corpora. They eventually use this information as features in a false friend classifier and reach up to 20% and 37% improvement over the baseline precision and recall respectively. Our false friend identification method is different from the mentioned studies in the sense that we generate a supervised classifier from fully unsupervised labeled data. Unlike previous work that solely focus on the identification task, our model leverages both identification and disambiguation.

From the sense disambiguation perspective, there have been several attempts to integrate word sense disambiguation (WSD) systems into the SMT framework in recent years. The main goal of these studies is to improve the target translation for an ambiguous word in the source sentence. Most studies in this area incorporate supervised WSD systems which exploit labeled training data. As an instance, Carpuat and Wu (Carpuat and Wu, 2005) integrate a supervised WSD model trained on the Senseval-3 Chinese lexical sample task data into a standard Chinese-English phrase-based SMT model with two methodologies: First, at the decode time, they limit set of translation candidates for an ambiguous word to the set of translations mapped to the sense predicted by the WSD model. Second, they replace the translations chosen by SMT with the translation predicted by WSD system. Nevertheless, they show none of these methods improves baseline BLEU score (Papineni et al., 2002). Vickrey et al (2005) formulate the task of using WSD for SMT as *word translation* task. They use parallel data to train their WSD model. They showed that they improve accuracy in both word translation and blank-filling tasks. However, they did not incorporate their word translation setup in an end-to-end SMT system.

Carpuat and Wu (2007) transformed the problem into a phrase sense disambiguation task by incorporating state-of-the-art WSD features for selecting a target phrase out of all aligned phrases as the possible senses. Chan et al (2007) also embedded state-of-the-art WSD system into SMT by adding more features into the SMT model. They showed that they improve Baseline BLEU score using their WSD-based model.

Yang and Kirchhoff (2012) use an unsupervised

WSD to improve SMT final performance. Similar to previous studies, they add the WSD acquired feature to the SMT model. They could improve the BLEU score by 0.3% compared to the baseline.

All the mentioned studies aim to enhance SMT by identifying the appropriate target translation for a source word in a given context. Our approach is different from previous work in two aspects: First, we try to improve SMT lexical choice by identifying false friends and replacing them with the most adequate equivalent from standard language. Unlike previous work, all these steps are done on a given input sentence and we can see them as a preprocessing phase, thereby, there is no need to change the SMT model. Second, our approach does not assume that the in-domain parallel data is available. Hence, it is not constrained by the domain and can be extended to any other language variants.

The main difference between this approach and our previous work as described in (Aminian et al., 2014) lies in the fact that we try to improve SMT lexical choice by enhancing FF translation. Rather than blindly replacing all dialectal words with their standard equivalent as we did in (Aminian et al., 2014), here we try to automatically identify FF as one of the important sources of translation degradation across language variants and leverage knowledge acquired from monolingual standard data to predict the best equivalent for FF based on the context.

3 Approach

We describe our model in this section. We use two modules in our model: 1) a FF identifier (henceforth PARL) and, 2) a disambiguator (henceforth WC). PARL is based on a supervised classifier. The training data for PARL is automatically obtained from parallel data. WC is based on the likelihood of each standard equivalent given the contextual information. In all of our definitions, we use DA and ST to refer to a dialectal and standard language, respectively.

3.1 PARL Classifier

We first give some basic definitions about the setup. Parallel text $\mathcal{D}$ is a set of aligned sentences $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_N\}$ and $\mathcal{S}' = \{\mathcal{S}'_1, \mathcal{S}'_2, ..., \mathcal{S}'_N\}$ in the source and target languages respectively. We assume $\mathcal{S}'$ to be English (EN) in our experiments. $\mathcal{S}$ contains both ST and DA sentences. Each training instance is shown with a tuple (k, i, y) in which $1 \leq k \leq |\mathcal{S}|$, $1 \leq i \leq |\mathcal{S}_k|$ and $y \in \mathcal{Y}$. $\mathcal{Y}$ refers to the set of labels {FF, NFF}. We represent the ith word in the kth sentence as w_{ki} and its features as $\phi(k, i) \in \mathbb{R}^d$ where d is the size of feature vector. Given the set of training tuples (k, i, y), a classification algorithm is used to train the model. We use Averaged Perceptron (Freund and Schapire, 1999) as our classifier with the following features: word form for the current word and part of speech tag for the previous, current and next words.

Automatic Label Estimation We use a dialect identification tool to define a function $\mathcal{L}(k, i)$, that identifies the dialect for w_{ki} out of two possibilities: DA and ST. We do word alignment on $\mathcal{D}$ using an unsupervised alignment algorithm. We define A_{ki} to be the English word aligned to w_{ki}. Accordingly, we define E^{ST}_w as the set of all English words aligned to the source word w for the cases where w is identified as ST. E^{ST}_w can be written as:

$$E^{ST}_w = \{\forall e \in EN |\, \exists j, h \;\; A_{j,h} = e,\, w_{j,h} = w,\, \mathcal{L}(j, h) = ST\} \quad (1)$$

To reduce noise in the automatically acquired word alignments, we just consider aligned word pairs with frequency more than 5. For each w_{ki} where $\mathcal{L}(k, i)$ is equal to DA, we have to decide whether the word is FF or not. We define a function $\mathcal{F}(w_{ki}, A_{ki})$ that returns true if we decide to label w_{ki} as FF and false otherwise (Eq. 2).

$$\mathcal{F}(w_{ki}, A_{ki}) = true \Leftrightarrow Sim(A_{ki}, E^{ST}_{w_{ki}}) < \delta \quad (2)$$

where δ is a manually defined threshold and Sim is defined in Eq. 3:

$$Sim(e, E) = \frac{1}{|\mathcal{C}_E|} \sum_{c \in \mathcal{C}_E} \frac{\sum_{e' \in c} dist(e, e')}{|c|} \quad (3)$$

where $\mathcal{C}_E$ partitions E into non-overlapping clusters. Each $c \in \mathcal{C}_E$ contains a cluster of words in E with similar meaning. The clusters are obtained from using the distance measure (Wu and Palmer, 1994) in Eq. 4.

$$dist(e, e') = \frac{2 \cdot d(s_{e,e'})}{d(e) + d(e')} \quad (4)$$

where $s_{e,e'}$ is a maximally specific superclass of e and e' in WordNet (Miller, 1995) and d is the depth of the node in the WordNet taxonomy.

In short, Eq. 3 computes a weighted average similarity between various ST senses of the target word and its DA sense in the sentence k. The intuition behind this setting is as follows: for a particular word that is identified as DA in a sentence, we measure similarity of its aligned English word to the set of all English words aligned to ST occurrences of the same word (E_w^{ST}). If this similarity is less than a threshold δ, we label that word as FF. We set δ to 0.5 in our experiments.

3.2 WC Classifier

We now describe our disambiguation model. We use a large amount of monolingual data $\mathcal{D}'$ as a set of sentences $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_M\}$ in the ST form. We perform unsupervised word clustering on $\mathcal{D}'$ to obtain word cluster assignments for each word. We then use word clusters to build our disambiguation model. The model comprises five parameters: $P_{-2}(c|w)$, $P_{-1}(c|w)$, $P_{+1}(c|w)$ and $P_{+2}(c|w)$ for all $c \in \{1, 2, ..., K\}$ where K refers to the number of clusters, in addition to the word probability $P(w)$. Hence, context parameters are $P_\tau(c|w)$ for $\tau \in \{-2, -1, +1, +2\}$ in which c specifies cluster of the word which is placed in the offset τ for the word w. $P_\tau(c|w)$ is estimated using maximum likelihood estimation with additive smoothing. The smoothing parameter is set to 0.1 in our experiments. To avoid sparsity, we assume that all previous contexts are the same and analogously all next contexts are also the same. In other words, we tie $P_{-2}(c|w)$ and $P_{-1}(c|w)$ into one parameter and $P_2(c|w)$ and $P_1(c|w)$ into another distinct parameter.

Let $\Omega(w)$ be the list of ST equivalents for the DA word w. We choose the most probable candidate ω^* using Eq. 5 by having $\tau \in \{-2, -1, 1, 2\}$.

$$\omega^* = \underset{\omega \in \Omega}{\operatorname{argmax}} \ \log P(\omega) + \sum_{\tau} \log P_\tau(c_\tau|\omega) \quad (5)$$

The intuition behind this model is as follows: if a particular DA word in a sentence is identified as FF, we want to replace it by one of its ST equivalents. If an alternative word is more likely to appear in that context compared to other possible equivalents, we expect our model to select that as the replacement. Since we train word clusters on ST data, the model tends to assign more weight on words that fit better to ST contexts.

4 Experimental Setup

Data Sets To train PARL classifier, we use parallel data $\mathcal{D}_{ME}$ which is a collection of MSA and EGY texts created from multiple LDC catalogs.[1] The data comprises 29M MSA and 5M DA tokenized words from multiple genres including newswire, broadcast news, broadcast conversations, and weblogs. To train the disambiguator, we use the Arabic Gigaword 4 (Graff and Cieri, 2003) containing 848M tokenized MSA words. To train the model described in § 3.2, we exclude punctuation as well as clitics from the target word local context. These words usually do not provide much information about the target word and will increase model sparsity. All data sets used in our experiments have undergone the following preprocessing steps: all Arabic data is Alef/Ya normalized and tokenized using MADAMIRA v1. (Pasha et al., 2014) according to Arabic Treebank (ATB) tokenization scheme (Maamouri et al., 2004). We used Tree Tagger (Schmid, 1994) to tokenize English data.

Tools We use GIZA++ (Och and Ney, 2003) for word alignment. We obtain word clusters from word2vec (Mikolov et al., 2013) K-means word clustering tool. We use the continuous bag of word model to build word vectors of size 200 using a word window of size 8 for both left and right. The number of negative samples for logistic regression is set to 25 and threshold used for sub-sampling of frequent words is set to 10^{-5} in the model with 15 iterations. We also use full softmax to obtain the probability distribution.

We use AIDA (Elfardy and Diab, 2013) as the dialect identification tool. AIDA also provides a list of MSA equivalents for identified DA words in context.

[1] 41 LDC catalogs including data prepared for GALE and BOLT projects.

	BLEU	METEOR	TER	WER	PER
BASELINE1	20.6	27.5	65.9	69.2	45.3
BASELINE2	20.1	27.2	68.3	71.6	46.6
BASELINE3	21.3	28.0	65.2	68.6	44.6
PARL	20.7	27.1	67.5	69.6	45.5
WC$_{cor}$	20.9	27.7	65.4	68.7	44.8
PARL+WC	21.0	27.7	66.2	68.5	45.3
PARL+WC$_{cor}$	21.3	27.9	65.5	**68.0**	**44.5**

Table 1: Evaluation results (BLEU, METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), WER, PER) on the Bolt-arz test set compared to the baselines.

SMT System We use Moses decoder (Koehn et al., 2007) to build a standard phrase-based SMT system. Feature weights are tuned to maximize BLEU score on the tuning set using Minimum Error Rate Training (MERT) (Och, 2003) algorithm. Final results are reported by averaging over three tuning sessions with random initialization. Significant test is also performed to make sure that gains in the results are statistically significant. We use the implementation of Clark et al. (2011) to compute the p-value via approximate randomization algorithms. Since AIDA generates MSA equivalents in the lemma form, we use a factored translation model with lemma and POS factors. We use GIZA++ (Och and Ney, 2003) to word align the parallel corpus. We use SRILM (Stolcke and others, 2002) to build 5-gram language models with modified Kneser-Ney smoothing (Kneser and Ney, 1995). Our language modeling data consists of three data sets: a) The English Gigaword 5 (Graff and Cieri, 2003); b) The English side of the BOLT Phase 1 parallel data; and, c) different LDC English corpora collected from discussion forums.[2].

The translation model is trained using the MSA part of $\mathcal{D}_{ME}$ with 29M words. Therefore, any improvement in translating DA words on the test set is gained by our false friend identification and disambiguation approach. Our test set comprises 16K tokenized EGY words and is acquired by selecting 1065 sentences from LDC2012E30 (BOLT-arz-test). The tuning set contains 1547 sentences obtained from multiple LDC catalogs[3] and comprises 20k tokens.

5 Results and Discussion

The main goal of this work is to improve the translation chosen by SMT for a false friend based on its surrounding context. The final SMT performance is affected by two factors: First, the accuracy of false friend identifier and disambiguator. Second, the quality of predefined candidates generated by AIDA (FF are then replaced by one of these candidates chosen by WC).

In order to accurately evaluate the quality of our identification and disambiguation process, we design three different baselines. As the first baseline, we randomly tag EGY words which have been observed as MSA in the train data as false friend. False friends then are replaced by a randomly selected sense from respective candidates list (BASELINE1). As the second baseline, we follow the setup introduced in (Aminian et al., 2014). In this baseline, all EGY words that meet mentioned criteria, are replaced with one randomly selected sense from the list of candidates (BASELINE2). As the third baseline, we use the results of the raw baseline without any replacement (BASELINE3). The first two baselines can be used to evaluate the accuracy of FF identifier and disambiguator modules. The last baseline evaluates the overall effectiveness of the approach to enhance EGY-EN SMT which depends on both factors mentioned before.

The first three rows of Table 1 show BASELINE1, BASELINE2 and BASELINE3 results on our test set. PARL in the fourth row demonstrates the setup that only parallel data is exploited to identify false friends. The identified DA word is then replaced by a randomly selected MSA sense from the candidate

[2] LDC2012E04, LDC2012E16, LDC2012E21, LDC2012E54

[3] LDC2012E15, LDC2012E19, LDC2012E55

Ref.	i will tell you a story , and you judge whose fault it is .
Baseline	**Tb** AnA H+ AHky l+ HDrp +k mwqf w+ tqwly myn Ally glTAn
Replacement	**tmAm** AnA H+ AHky l+ HDrp +k mwqf w+ tqwly myn Ally glTAn
Baseline Trans.	ok , i am going to talk to you and say who was wrong .
Replacement Trans.	i will talk to you stand and say who was wrong .

Table 2: Example of correct FF identification and replacement with non-improving BLEU score.

list. Similarly, WC$_{cor}$ shows the setup where WC is directly used to identify and replace false friends. In this setup, original EGY word is manually added to the list of MSA candidates generated by AIDA. Thus, WC module selects the most adequate candidate based on the context from the list containing both MSA equivalents and original EGY word. In other words, WC simultaneously performs FF identification and sense disambiguation. PARL+WC refers to the system that uses PARL to identify FF and then WC to disambiguate them. It is to be emphasized that in this setup, WC chooses the most appropriate MSA equivalent of each false friend only from the list of candidates generated by AIDA. We also define PARL+WC$_{cor}$ in which WC$_{cor}$ is used as a FF identifier as well as disambiguator (similar to the second setup above). In fact, we prevent mistakes from PARL by using WC as an identifier as well. This setup replaces a word by its MSA equivalent only if both PARL and WC identify it as FF.

As shown in Table 1, all replacement experiments outperform BASELINE1 and BASELINE2 in terms of BLEU score. PARL improves BASELINE1 and BASELINE2 BLEU scores by 0.1% absolute (0.5% relative) and 0.6% absolute (3% relative) receptively. This implies that our FF identifier achieves more accurate FF predictions compared to random and blind predictions.

Using WC$_{cor}$ for FF identification and disambiguation shows a noticeable improvement over the case that we just use PARL for identification (in terms of BLEU, WER and PER). This shows that contextual similarity plays a more important role compared to the information extracted from parallel data to train a FF identification model. PARL is also too sensitive to errors in the word alignment. So noise in the alignment will lead to incorrect prediction and thereby, inadequate replacement.

As expected, combining PARL and WC for FF identification and replacement (PARL+WC) outperforms the individual decisions made by each module solely. This setup benefits from evidences provided by both modules for FF identification and sense disambiguation. Eventually, the last setup PARL+WC$_{cor}$ leads to 0.3% absolute (1.4% relative) BLEU improvement over PARL+WC. It also outperforms PARL+WC in terms of other SMT evaluation metrics such as METEOR, TER, WER and PER. For example, it achieves 0.7%, 0.5% and 0.8% reduction in TER, WER and PER respectively compared to PARL+WC. In the last setup, we just replace words which both PARL and WC commonly identify them as FF. In other words, WC refines some of the PARL mistakes and avoids it from replacing words which are mistakenly identified as FF by PARL . It is worth noting that significant tests show that all gains in the BLEU, METEOR and TER over BASELINE2 and BASELINE3 are statistically significant at the 95% level.

Our best performing setup, PARL+WC$_{cor}$, reduces BASELINE3 WER and PER by the noticeable amount of 0.6% and 0.1% respectively. This indicates that our approach has the power to enhance SMT lexical choice and select more accurate target translations for the false friends. However, our method does not outperform BASELINE3 BLEU score. Our analysis shows that the main reason for this phenomenon is that the SMT translation table does not contain adequate bilingual phrase pairs for some of the replaced MSA equivalents (suggested by AIDA). Thus, decoder can not generate coherent phrases while translating these words. As an example, consider the sentence shown in Table 2. Word ‘*Tb*’ in the baseline sentence means *all right*, *very well* or *ok* in EGY while it means *medicine* when used as MSA. Our FF identifier has correctly identified this word as a FF. The disambiguator module also has adequately replaced word ‘Tb’ with the MSA word ‘tmAm’ which means *ok*. However, this replacement does not yield to a better translation for this word. This happens because word 'tmAm' has not been observed as an interjection in our SMT phrase table. Thus, SMT decoder is not able to find a good translation for this word.

	BASELINE1	BASELINE2	BASELINE3
PARL	37.7/38.5	41.5/40.3	34.7/44.2
PARL+WC	38.7/32.8	45.5/36.4	34.0/35.2
PARL+WC$_{cor}$	40.5/32.2	46.4/36.3	35.7/35.1

Table 3: Percentage of BLEU-enhanced sentences/percentage of BLEU-degraded sentences for different replacement approaches compared to each baseline separately.

We conducted another analysis to closely assess the impact of our disambiguator module (WC) in improving target sentences BLEU score. We ran our replacement setups on the proportion of Bolt-arz sentences which contain at least one FF. FF are predicted by PARL module. We ended up getting a set with 796 sentences. Table 3 shows the percentage of BLEU-enhanced and BLEU-degraded sentences in this set for each setup compared to the baselines separately. The setup which exploits WC$_{cor}$ for FF identification and disambiguation is excluded from this comparison as it does not use PARL for FF identification. As the percentages in Table 3 indicate, PARL+WC noticeably increases (decreases) percentage of BLEU-enhanced (BLEU-degraded) sentences compared to PARL setup with respect to BASELINE1 and BASELINE2. As shown before (Table 1), the last setup PARL+WC$_{cor}$ did not improve BASELINE3 BLEU score. However, results in Table 3 show that this setup increases percentage of BLEU-enhanced sentences compared to PARL+WC and PARL with respect to BASELINE3 significantly. Comparing percentages of BLEU˙degraded sentences for mentioned setups gives the same results.

Table 4 shows some translation examples with and without any replacement. The replacement is done using our best-performing setup PARL+WC$_{cor}$ on Bolt-arz test set. The first four examples demonstrate cases that FF (shown in bold) are correctly identified and replaced with a proper MSA equivalent. For instance, the word 'zy' in the first example means *uniform* or *clothing* in MSA and *such as* or *like* in EGY. Thus, replacing the word 'zy' with MSA word 'mvl' which means *like* yields to better translation and thereby, improves BLEU score.

In the second example, word 'nsyb' which means *forget* in this context is replaced with MSA equivalent 'trk' that means *leave* or *forget*. As the result, decoder has translated phrase 'trk +nA mn AlAxtlAf' into a longer phrase *let us from the difference* instead of generating an incoherent translation such as baseline.

Word 'wHcp' in the third example is not a pure EGY word. However, it conveys a meaning different from its observed senses in the phrase table. Hence, baseline incorrectly translates this word to *difficult* while the replaced setup generate the correct translation *bad* for the replaced MSA equivalent 'syC'. Hence, as shown, our approach has improved SMT lexical choice significantly in this example.

Word 'cwf' in the fourth example is also correctly identified as a FF according to context. This word is used as noun in MSA with meanings *look* and *appearance* while it is used as a command verb (*order someone to look*) in EGY. As we can see, our disambiguator module has adequately replaced this word with the verb 'rAy' which means *to look at* or *to see*. As the result, the decoder has translated this word into the word *see* in the English sentence which leads to higher BLEU score compared to the baseline translation.

Word 'Erkp' in the fifth example has English equivalent *battle* in EGY and *test* in MSA context. Similar to the previous example, baseline selects the incorrect translation *testing*. While our replacement setup substitues this word with MSA equivalent 'mErkp' which means *battle* and thereby, improves the translation.

Sixth instance in Table 2 demonstrates the example where our FF identifier has incorrectly identified word 'HAjp' (*need* in this context) as FF. This word is then replaced by the word 'Amr' (*order*) which does not convey the original word meaning according to context. Hence, the decoder is not able to find a proper translation for the replaced word in the context.

Ref.	not private , i mean like buses and the metro and trains ... etc .
Baseline	mc mlkyp xASp yEny AqSd **zy** AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx
Replacement	mc mlkyp xASp yEny AqSd **mvl** AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx
Baseline Trans.	privately , i mean , i mean , i do not like the bus and subway train , etc .
Replacement Trans.	not privately , i mean , i mean , such as the bus and subway train , etc .
Ref.	let us forget about our differences and unite .
Baseline	**nsyb** +nA mn AlAxtlAf w+ ntwHd
Replacement	**trk** +nA mn AlAxtlAf w+ ntwHd
Baseline Trans.	we disagree and suffering from
Replacement Trans.	let us from the difference and unify
Ref.	and those who said that the girls ... indeed , i heard very bad words , why ?
Baseline	w+ Ally yqwl AlbnAt . . . b+ jd smEt AllfAZ **wHcp** qwy lyh kdh
Replacement	w+ Ally yqwl AlbnAt . . . b+ jd smEt AllfAZ **syC** qwy lyh kdh
Baseline Trans.	and to say ... very very difficult . that is why i heard
Replacement Trans.	and to say ... seriously , i heard a strong bad , why ?
Ref.	at least three parties ; check them and read about them in detail
Baseline	Ely AlAql three AHzAb **cwf** +hm w+ AqrA +hm b+ Emq
Replacement	Ely AlAql three AHzAb **rAy** +hm w+ AqrA +hm b+ Emq
Baseline Trans.	at least three of the depth of them and with them .
Replacement Trans.	at least three parties see them and baqir them in depth
Ref.	it is waiting for disagreement between the salafis and the liberals , which engages them in a new battle of nonsense speech similar to
Baseline	yntZr An yxtlf Alslfywn mE AllybrAlyyn f+ ydxlwA fy **Erkp** Ely +k jdydp mn qbyl rmy
Replacement	yntZr An yxtlf Alslfywn mE AllybrAlyyn f+ ydxlwA fy **mErkp** Ely +k jdydp mn qbyl rmy
Baseline Trans.	it is expected that the salafis disagrees with liberals , in testing on your new prior to throw
Replacement Trans.	waiting for the salafis disagrees with liberals , in the battle for your new prior to throw
Ref.	also eradication of poverty and need is very important , toqua
Baseline	w+ kmAn AlqDAC Ely Alfqr w+ **HAjp** mhm jdA yA+ tqy
Replacement	w+ kmAn AlqDAC Ely Alfqr w+ **Amr** kbyr jdA yA+ tqy
Baseline Trans.	and also the eradication of poverty and need is very important ,
Replacement Trans.	and also the eradication of poverty and a very large ,

Table 4: Translation examples with and without replacement drawn from Bolt-arz test

6 Conclusion and Future Work

We presented a new approach for improving cross-language SMT performance without any in-domain training data by identifying false friends and replacing them with a semantically similar equivalent from the standard language. We show that our approach improves lexical choice in EGY-EN SMT system trained only on MSA data. We demonstrate a fully unsupervised approach for false friend identification and disambiguation using evidences extracted from parallel and monolingual data. We showed that our best-performing setup reduces the baseline WER and PER by the noticeable amount of 0.6% and 0.1% respectively. One interesting line to expand this study is exploring an automatic way to generate the list of possible equivalents for FF instead of using a predefined inventory of senses. One idea is benefiting from continues word vectors and their similarity to extract possible word senses for a particular FF from available monolingual corpus.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) Contract No. HR0011-12-C-0014, the BOLT program with subcontract from Raytheon BBN. We would like to acknowledge the useful comments by three anonymous reviewers who helped in making this publication more concise and better presented.

References

Maryam Aminian, Mahmoud Ghoneim, and Mona Diab. 2014. Handling oov words in dialectal arabic to english machine translation. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 99–108, Doha, Qatar, October. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Keith Brown and Keith Allan. 2010. *Concise encyclopedia of semantics*. Elsevier.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June. Association for Computational Linguistics.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

Oana Frunza and Diana Inkpen. 2006. Semi-supervised learning of partial cognates using bilingual bootstrapping. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 441–448, Sydney, Australia, July. Association for Computational Linguistics.

David Graff and Christopher Cieri. 2003. English gigaword, ldc catalog no.: Ldc2003t05. *LDC2003T05. Linguistic Data Consortium, University of Pennsylvania*.

Nizar Y Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

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 the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, pages 102–109.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine translation*, 21(1):29–53.

Andrea Mulloni, V Pekar, R Mitkov, and D Blagoev. 2007. Semantic evidence for automatic identification of cognates. In *Proceedings of the RANLP2007 workshop: Acquisition and management of multilingual lexicons*, pages 49–54. Citeseer.

Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2007. Cognate or false friend? ask the web. In *Proceedings of the RANLP2007 workshop: Acquisition and management of multilingual lexicons*, pages 55–62.

Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2009. Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of the International Conference RANLP-2009*, pages 292–298, Borovets, Bulgaria, September. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1479.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Stefan Schulz, Kornel Markó, Eduardo Sbrissia, Percy Nohama, and Udo Hahn. 2004. Cognate mapping - a heuristic strategy for the semi-supervised acquisition of a spanish lexicon from a portuguese seed lexicon. In *Proceedings of Coling 2004*, pages 813–819, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Andreas Stolcke et al. 2002. Srilm an extensible language modeling toolkit. In *INTERSPEECH*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proceedings of Eurospeech'97*, pages 2667–2670.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 771–778, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA, June. Association for Computational Linguistics.

Mei Yang and Katrin Kirchhoff. 2012. Unsupervised translation disambiguation for cross-domain statistical machine translation. In *Proceedings of association for machine translation in the Americas*.