Handling OOV Words in Dialectal Arabic to English Machine Translation

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

Abstract

Dialects and standard forms of a language typically share a set of cognates that could bear the same meaning in both varieties or only be shared homographs but serve as *faux amis*. Moreover, there are words that are used exclusively in the dialect or the standard variety. Both phenomena, faux amis and exclusive vocabulary, are considered out of vocabulary (OOV) phenomena. In this paper, we present this problem of OOV in the context of machine translation. We present a new approach for dialect to English Statistical Machine Translation (SMT) enhancement based on normalizing dialectal language into standard form to provide equivalents to address both aspects of the OOV problem posited by dialectal language use. We specifically focus on Arabic to English SMT. We use two publicly available dialect identification tools: AIDA and MADAMIRA, to identify and replace dialectal Arabic OOV words with their modern standard Arabic (MSA) equivalents. The results of evaluation on two blind test sets show that using AIDA to identify and replace MSA equivalents enhances translation results by 0.4% absolute BLEU (1.6% relative BLEU) and using MADAMIRA achieves 0.3% absolute BLEU (1.2% relative BLEU) enhancement over the baseline. We show our replacement scheme reaches a noticeable enhancement in SMT performance for faux amis words.

1 Introduction

In this day of hyper connectivity, spoken vernaculars are ubiquitously ever more present in textual social media and informal communication channels. Written (very close to the spoken) informal language as represented by dialect poses a significant challenge to current natural language processing (NLP) technology in general due to the lack of standards for writing in these vernaculars. The problem is exacerbated when the vernacular constitutes a dialect of the language that is quite distinct and divergent from a language standard and people code switch within utterance between the standard and the dialect. This is the case for Arabic. Modern Standard Arabic (MSA), as the name indicates, is the official standard for the Arabic language usually used in formal settings, while its vernaculars vary from it significantly forming dialects known as dialectal Arabic (DA), commonly used in informal settings such as the web and social media. Contemporary Arabic is a collection of these varieties. Unlike MSA, DA has no standard orthography (Salloum and Habash, 2013). Most of the studies in Arabic NLP have been conducted on MSA. NLP research on DA, the unstandardized spoken variety of Arabic, is still at its infancy. This constitutes a problem for Arabic processing in general due to the ubiquity of DA usage in written social media. Moreover, linguistic code switching between MSA and DA always happens either in the course of a single sentence or across different sentences. However this intrasentential code switching is quite pervasive (Elfardy et al., 2013). For instance 98.13% of sentences crawled from Egyptian DA (EGY) discussion forums for the COLABA project (Diab et al., 2010) contains intrasentential code switching.

MSA has a wealth of NLP tools and resources compared to a stark deficiency in such resources for DA. The mix of MSA and DA in utterances constitutes a significant problem of Out of Vocabulary (OOV) words in the input to NLP applications. The OOV problem is two fold: completely unseen words in training data, and homograph OOVs where the word appears in the training data but with a different sense. Given these issues, DA

Language Technology for Closely Related Languages and Language Variants (LT4CloseLang), pages 99–108,
October 29, 2014, Doha, Qatar.

NLP and especially DA statistical machine translation (SMT) can be seen as highly challenging tasks and this illustrates the need for conducting more research on DA.

MSA has a wealth of resources such as parallel corpora and tools like morphological analyzers, disambiguation systems, etc. On the other hand, DA still lacks such tools and resources. As an example, parallel DA to English (EN) corpora are still very few and there are almost no MSA-DA parallel corpora. Similar to MSA, DA has the problem of writing with optional diacritics. It also lacks orthographic standards. Hence, translating from DA to EN is challenging as there are impediments posed by the nature of the language coupled with the lack of resources and tools to process DA (Salloum and Habash, 2013).

MSA and DA are significantly different on all levels of linguistic representation: phonologically, morphologically, lexically, syntactically, semantically and pragmatically. The morphological differences between MSA and DA are most noticeably expressed by using some clitics and affixes that do not exist in MSA. For instance, the DA (Egyptian and Levantine) future marker clitic *H*[1] is expressed as the clitic *s* in MSA (Salloum and Habash, 2013). On a lexical level, MSA and DA share a considerable number of faux amis where the lexical tokens are homographs but have different meanings. For instance the word *yEny* in MSA means 'to mean', but in DA, it is a pragmatic marker meaning 'to some extent'. We refer to this phenomenon as sense OOV (SOOV). This phenomenon is in addition to the complete OOV (COOV) that exist in DA but don't exist in MSA. These issues constitute a significant problem for processing DA using MSA trained tools. This problem is very pronounced in machine translation.

In this paper, we present a new approach to build a DA-to-EN MT system by normalizing DA words into MSA. We focus our investigation on the Egyptian variety of DA (EGY). We leverage MSA resources with robust DA identification tools to improve SMT performance for DA-to-EN SMT. We focus our efforts on replacing identified DA words by MSA counterparts. We investigate the replacement specifically in the decoding phase of the SMT pipeline. We explore two state of the art DA identification tools for the purposes of our study. We demonstrate the effects of our replacement scheme on each OOV type and show that normalizing DA words into their equivalent MSA considerably enhances SMT performance in translating SOOVs.

The remainder of this paper is organized as follows: Section 2 overviews related work; Section 3 details our approach; Section 4 presents the results obtained on standard data sets; in Section 5, we discuss the results and perform error analysis; finally we conclude with some further observations in Section 6.

2 Related Work

Leveraging MSA resources and tools to enrich DA for NLP purposes has been explored in several studies. Chiang, et. al. (2006) exploit the relation between Levantine Arabic (LEV) and MSA to build a syntactic parser on transcribed spoken LEV without using any annotated LEV corpora.

Since there are no DA-to-MSA parallel corpora, rule-based methods have been predominantly employed to translate DA-to-MSA. For instance, Abo Bakr et al. (2008) introduces a hybrid approach to transfer a sentence from EGY into a diacritized MSA form. They use a statistical approach for tokenizing and tagging in addition to a rule-based system for constructing diacritized MSA sentences. Moreover, Al-Sabbagh and Girju (2010) introduce an approach to build a DA-to-MSA lexicon through mining the web.

In the context of DA translation, Sawaf (2010) introduced a hybrid MT system that uses statistical and rule-based approaches for DA-to-EN MT. In his study, DA words are normalized to the equivalent MSA using a dialectal morphological analyzer. This approach achieves 2% absolute BLEU enhancement for Web texts and about 1% absolute BLEU improvement over the broadcast transmissions. Furthermore, Salloum and Habash (2012) use a DA morphological analyzer (ADAM) and a list of hand-written morphosyntactic transfer rules (from DA to MSA) to improve DA-to-EN MT. This approach improves BLEU score on a blind test set by 0.56% absolute BLEU (1.5% relative) on the broadcast conversational and broadcast news data. Test sets used in their study contain a mix of Arabic dialects but Levantine Arabic constitutes the majority variety.

Zbib et al. (2012) demonstrate an approach to ac-

[1]We use the Buckwalter Transliteration as represented in www.qamus.com for Romanized Arabic representation throughout the paper.

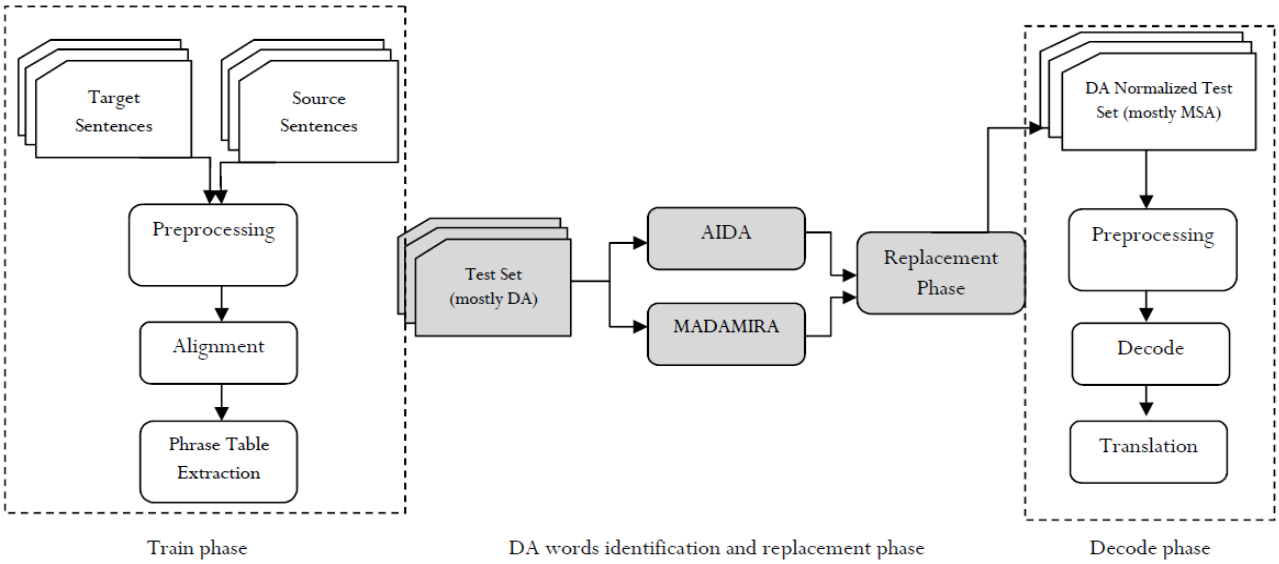

Figure 1: Block diagram of the proposed system for enhancing DA-to-EN SMT via normalizing DA

quire more DA-to-EN data to improve DA SMT performance by enriching translation models with more DA data. They use Amazon Mechanical Turk to create a DA-to-EN parallel corpus. This parallel data is augmented to the available large MSA-to-EN data and is used to train the SMT system. They showed that their trained SMT model on this DA-to-EN data, can achieve 6.3% and 7% absolute BLEU enhancement over an SMT system trained on MSA-to-EN data when translating EGY and LEV test sets respectively. Habash (2008) demonstrates four techniques for handling OOV words through modifying phrase tables for MSA. He also introduces a tool which employs these four techniques for online handling of OOV in SMT (Habash, 2009).

Habash et al. (2013) introduces MADA-ARZ, a new system for morphological analysis and disambiguation of EGY based on an MSA morphological analyzer MADA (Habash and Rambow, 2005). They evaluate MADA-ARZ extrinsically in the context of DA-to-EN MT and show that using MADA-ARZ for tokenization leads to 0.8% absolute BLEU improvement over the baseline which is simply tokenized with MADA. In this paper, we use MADAMIRA (Pasha et al., 2014), a system for morphological analysis and disambiguation for both MSA and DA (EGY), to identify DA words and replace MSA equivalents. Our approach achieves 0.6% absolute BLEU improvement over the scores reported in (Habash et al., 2013).

3 Approach

In the context of SMT for DA-to-EN, we encounter a significant OOV rate between test and training data since the size of the training data is relatively small. On the other hand, we have significant amounts of MSA-to-EN parallel data to construct rich phrase tables. MSA and DA, though divergent, they share many phenomena that can be leveraged for the purposes of MT. Hence, if we combine training data from MSA with that from DA, and then at the decode time normalize OOV DA words into their equivalent MSA counterparts we should be able to overcome the resource challenges in the DA-to-EN SMT context, yielding better overall translation performance. The OOV problem is two fold: complete OOV (COOV) and sense OOV (SOOV). The COOV problem is the standard OOV problem where an OOV in the input data is not attested at all in the training data. The SOOV problem is where a word is observed in the training data but with a different usage or sense, different from that of the test data occurrence. To our knowledge, our research is the first to address the SOOV directly in the context of SMT. To that end, we employ two DA identification tools: a morphological tagger, as well as a full-fledged DA identification tool to identify and replace DA words with their equivalent MSA lemmas in the test data at decoding time.

Accordingly, the ultimate goal of this work is to assess the impact of different DA identification and replacement schemes on SMT overall perfor-

mance and overall OOV (both types) reduction. It is worth noting that we focus our experiments on the decoding phase of the SMT system. Figure1 shows the block diagram of the proposed system. We exploit the following tools and resources:

- **MADAMIRA:** A system for morphological analysis and disambiguation for both MSA and DA (EGY). MADAMIRA indicates whether a word is EGY or MSA based on its underlying lexicon which is used to generate an equivalent EN gloss. However, for EGY words, MADAMIRA does not generate the equivalent MSA lemma (Pasha et al., 2014);
- **AIDA:** A full-fledged DA identification tool which is able to identify and classify DA words on the token and sentence levels. AIDA exploits MADAMIRA internally in addition to more information from context to identify DA words (Elfardy and Diab, 2013). AIDA provides both the MSA equivalent lemma(s) and corresponding EN gloss(es) for the identified DA words;
- **THARWA:** A three-way lexicon between EGY, MSA and EN (Diab et al., 2014).

To evaluate effectiveness of using each of these resources in OOV reduction, we have exploited the following replacement schemes:

- AIDA identifies DA words in the context and replaces them with the most probable equivalent MSA lemma;
- MADAMIRA determines whether a word is DA or not. If the word is DA, then EN gloss(es) from MADAMIRA are used to find the most probable equivalent MSA lemma(s) from THARWA.

As all of these DA identification resources (MADAMIRA, AIDA and THARWA) return MSA equivalents in the lemma form, we adopt a factored translation model to introduce the extra information in the form of lemma factors. Therefore, DA replacement affects only the lemma factor in the factored input. We consider the following setups to properly translate replaced MSA lemma to the the corresponding inflected form (lexeme):[2]

- Generated lexeme-to-lexeme translation (Glex-to-lex): To derive inflected MSA lexeme from MSA replaced lemma and POS, we construct a generation table on the factored data to map lemma and POS factors into lexeme. This table is generated using Moses toolkit (Koehn et al., 2007) generation scripts and provides a list of generated lexemes for each lemma-POS pair. An MSA lexeme language model (LM) is then used to decode the most probable sequence of MSA lexemes given these generated lexemes for each word in the sentence.
- lemma+POS-to-lexeme translation (lem+POS-to-lex): In this path source lemma and POS are translated into the appropriate target lexeme. We expect this path provides plausible translations for DA words that are not observed in the phrase tables.
- lexeme-to-lexeme;lemma+POS-to-lexeme translation (lex-to-lex;lem+POS-to-lex): The first path translates directly from a source lexeme to the target lexeme. So it provides appropriate lexeme translations for the words (MSA or DA) which have been observed in the trained model. It is worth noting that lex-to-lex translation path does not contain any replacement or normalization. Therefore, it is different from the first path (Glex-to-lex). The second path is similar to the lem+POS-to-lex path and is used to translate DA words that do not exist in the trained model.

3.1 Data Sets

For training translation models we use a collection of MSA and EGY texts created from multiple LDC catalogs[3] comprising multiple genres (newswire, broadcast news, broadcast conversations, newsgroups and weblogs).The train data contains 29M MSA and 5M DA tokenized words. We use two test sets to evaluate our method on both highly DA and MSA texts: For DA test data, we selected 1065 sentences from LDC2012E30, which comprises 16177 tokenized words (BOLT-arz-test); For MSA, we use the NIST MTEval 2009 test set (LDC2010T23), which contains

[2] We use the term lexeme to indicate an inflected tokenized uncliticized form of the lemma. A lemma in principle is a lexeme but it is also a citation form in a dictionary.

[3] 41 LDC catalogs including data prepared for GALE and BOLT projects. Please contact the authors for more details.

1445 sentences corresponding to 40858 tokenized words (MT09-test). As development set (dev set), we randomly select 1547 sentences from multiple LDC catalogs (LDC2012E15, LDC2012E19, LDC2012E55) which comprises 20780 tokens.
The following preprocessing steps are performed on the train, test and dev sets: The Arabic side of the parallel data is Alef/Ya normalized and tokenized using MADAMIRA v1. according to Arabic Treebank (ATB) tokenization scheme (Maamouri et al., 2004); Tokenization on the EN side of the parallel data is performed using Tree Tagger (Schmid, 1994).

3.2 Language Modeling

We create a 5-gram language model (LM) from three corpora sets: a) The English Gigaword 5 (Graff and Cieri, 2003); b) The English side of the BOLT Phase1 parallel data; and, c) different LDC English corpora collected from discussion forums (LDC2012E04, LDC2012E16, LDC2012E21, LDC2012E54). We use SRILM (Stolcke., 2002) to build 5-gram language models with modified Kneser-Ney smoothing.

3.3 SMT System

We use the open-source Moses toolkit (Koehn et al., 2007) to build a standard phrase-based SMT system which extracts up to 8 words phrases in the Moses phrase table. The parallel corpus is word-aligned using GIZA++ (Och and Ney, 2003). Feature weights are tuned to maximize BLEU on the dev set using Minimum Error Rate Training (MERT) (Och, 2003). To account for the instability of MERT, we run the tuning step three times per condition with different random seeds and use the optimized weights that give the median score on the development set. As all our DA identification resources (MADAMIRA, AIDA and THARWA) are lemma-based, we adopt a factored translation model setup to introduce the extra information in the form of a lemma factor. As lemma only is not enough to generate appropriate inflected surface (lexeme) forms, we add a POS factor with two main translation paths: (i) direct translation from a source lexeme to the target lexeme; and (ii) translation from source lemma and POS to the appropriate target lexeme. Therefore, the first path should provide plausible translations for the words that have been seen before in the phrase tables while we expect that the second path provides feasible translations for DA words that are not seen in the trained model.

4 Experimental Results

4.1 Baseline Results

For each experimental condition mentioned in Section 3, we define a separate baseline with similar setup. These baselines use the SMT setup described in Section 3.3 and are evaluated on the two test sets mentioned in Section 3.1. To assess effectiveness of normalizing DA into MSA on the overall performance of MT system, the dev and test sets are processed through the similar steps to generate factored data but without any replacement of the DA words with MSA correspondents. We believe this to be a rigorous and high baseline as data contains some morphological information useful for DA-to-EN translation in the form of lemma and POS factors. We started with a baseline trained on the 29M words tokenized MSA training set and 5M words tokenized DA set separately. We created the baseline trained on the 34M words MSA+DA train data. Our objective of splitting train data based on its dialectal variety is to assess the role of DA words existing in the train set in the performance of our approach.
Table 1 illustrates baseline BLEU scores on BOLT-arz and MT09-test sets with three different training conditions: MSA+DA, MSA only, and DA only.

4.2 Replacement Experimental Results

We run the SMT pipeline using the feature weights that performed best during the tuning session on our dev set. Then the SMT pipeline with these tuned weights is run on two blind test sets. To account for statistical significance tests we used bootstrapping methods as detailed in (Zhang and Vogel, 2010). Table 2 shows BLEU scores of different DA identification and replacement schemes exploited in different setups on the test sets.
As we can see in Table 2, both AIDA and MADAMIRA replacement schemes outperform the baseline scores using MSA+DA trained models and lem+POS-to-lex;lex-to-lex setup. AIDA reaches 0.4% absolute BLEU (1.6% relative BLEU) improvement and MADAMIRA achieves 0.3% absolute BLEU (1.2% relative BLEU) enhancement over the corresponding baselines. This is while the same enhancement in BLEU scores can not be captured when we exploit the model

Test Set	Train Set	lex-to-lex	lem+POS-to-lex	lex-to-lex:lem+POS-to-lex
BOLT-arz-test	MSA+DA	26.2	25.4	25.5
	MSA	21.8	21.2	21.8
	DA	24.3	24.6	24.8
MT09-test	MSA+DA	48.2	46.9	47.3
	MSA	44.4	45.4	44.6
	DA	35.6	36.1	34.2

Table 1: Baseline BLUE scores for each setup on two test sets: BOLT-arz-test and MT09-test. Results are reported for each training input language variety separately.

Test Set	Train Set	Glex-to-lex		lem+POS-to-lex		lex-to-lex:lem+POS-to-lex	
		AIDA	MADAMIRA	AIDA	MADAMIRA	AIDA	MADAMIRA
BOLT-arz-test	MSA+DA	24.4	25.1	22.6	24.1	**25.9**	25.8
	MSA	20.6	21.0	20.1	20.3	21.7	22.0
	DA	24.3	23.7	21.3	23.1	24.5	24.8
MT09-test	MSA+DA	45.9	45.8	45.4	44.6	47.1	47.3
	MSA	42.7	42.4	45.2	43.7	44.5	44.6
	DA	35.6	34.0	36.1	34.5	34.1	34.3

Table 2: BLEU scores of AIDA and MADAMIRA replacement for the different setups on BOLT-arz-test and MT09-test. Results are reported for each training language variety separately.

which is trained on MSA or DA parallel data solely. This indicates that normalizing DA into MSA can reach its best performance only when we enrich the training model with DA words at the same time. Therefore, we note that acquiring more DA data to enrich phrase tables at the training phase and normalizing DA at the decoding step of SMT system would yield the best DA-to-EN translation accuracy.

Regardless of the replacement scheme we use to reduce the OOV rate (AIDA or MADAMIRA), BLEU scores on the MT09 are much higher than those on the BOLT-arz because the amount of MSA words in the training data is much more than DA words. Therefore, SMT system encounters less OOVs at the decode time on MSA texts such as MT09. Overall we note that adding AIDA or MADAMIRA to the setup at best has no impact on performance on the MT09 data set since it is mostly MSA. However, we note a small impact for using the tools in the lex-to-lex:lem+POS-to-lex path in the MSA+DA experimental setting.

Comparing results of different setups indicates that adding lex-to-lex translation path to the lem+POS-to-lex increases both AIDA and MADAMIRA performance on two test sets significantly. As Table 2 demonstrates adding lex-to-lex path to the lem+POS-to-lex translation using the model trained on MSA+DA data leads to 3.3% and 1.7% BLEU improvement using AIDA and MADAMIRA, respectively on the BOLT-arz set. Similar conditions on the MT09-test gives us 1.7% and 0.7% absolute improvement in the BLEU scores using AIDA and MADAMIRA respectively. This happens because lex-to-lex path can provide better translations for the words (MSA or DA) which have been seen in the phrase tables and having both these paths enables the SMT system to generate more accurate translations. Our least results are obtained when we use lem+POS-to-lex translation path solely either using AIDA or MADAMIRA which mainly occurs due to some errors existing in the output of morphological analyzer that yields to the erroneous lemma or POS.

	BOLT-arz	MT09
Sent.	1065	1445
Types	4038	8740
Tokens	16177	40858
COOV (type)	126 (3%)	169 (2%)
COOV (token)	134 (0.82%)	187 (0.45%)

Table 3: Number of sentences, types, tokens and COOV percentages in each test set

Reference	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
Baseline translation	privately , i mean , i mean , i do not like the bus and metro and train , etc .
Replacement	mc mlkyp xASp yEny AqSd **mvl** AlAtwbys w+ Almtrw w+ AlqTAr . . . Alx
Replacement translation	not a private property , i mean , i mean , like the bus and metro and train , etc .

Table 4: Example of translation enhancement by SOOV replacement

5 Error Analysis

To assess the rate of OOV reduction using different replacement methodologies, we first identify OOV words in the test sets. Then, out of these words, cases that our approach has led to an improvement in the sentence BLEU score over the baseline is reported. Table 3 shows the number of sentences, types and tokens for each test set as well as the corresponding type and token OOV counts. As we can see in this table, 0.82% of tokens in BOLT-arz and 0.45% of tokens in MT09-test sets are OOV. These cover the complete OOV cases (COOV).

In addition to these cases of COOV that are caused by lack of enough training data coverage, there are sense OOV (SOOV). SOOV happens when a particular word appears in both DA and MSA data but have different senses as faux amis. For instance the Arabic word *qlb* occurs in both MSA and DA contexts but with a different set of senses due to the lack of diacritics. In the specific MSA context it means 'heart' while in DA it means either 'heart' or 'change'. Therefore, in addition to the cases that word sense is triggered by DA context, other levels of word sense ambiguity such as homonymy and polysemy are involved in defining an SOOV word. Hence, SOOV identification in the test set needs additional information such as word equivalent EN gloss.

We determine SOOV as the words that (i) are observed as MSA word in the training data and considered a DA word in the test set once processed by AIDA and MADAMIRA; and, (ii) MSA and DA renderings have different non-overlapped equivalent EN glosses as returned by our AIDA and MADAMIRA. We assume that words with different dialectal usages in the train and test will have completely different EN equivalents, and thereby will be considered as SOOV. One of the words that this constraint has recognized as SOOV is the word *zy* with English equivalent 'uniform' or 'clothing' in MSA and 'such as' or 'like' in DA. Replacement of this SOOV by the MSA equivalent 'mvl' yields better translation as shown in Table 4.

Among all COOV words, our approach only targets COOV which are identified as DA. Table 5 and 6 report the number of COOV words (type and token) which have been identified as DA by AIDA or MADAMIRA in BOLT-arz and MT09 test sets, respectively. Second column in these tables represent number of SOOV (type and token) in each set. Last columns show percentage of sentences which have had at least one COOV or SOOV word and our replacement methodology has improved the sentence BLEU score over the baseline for each setup, respectively. Percentages in these columns demonstrate the ratio of enhanced sentences to the total number of sentences which have been determined to have at least one COOV or SOOV word. These percentages are reported on the MSA+DA data to train the SMT system condition.

While Table 5 and 6 show enhancements through DA COOV replacements, our manual assessment finds that most of these enhancements are actually coming from SOOVs present in the same sentences. For example, when we examined the 21 types identified by AIDA as DA COOV in BOLT-arz we found 9 typos, 5 MSAs, one foreign word and only 6 valid DA types. Moreover, none of the replacements over these 6 DA types yield an enhancement.

Although Table 5 shows that MADAMIRA achieves more success enhancing BLEU score of sentences which contain SOOV words on BOLT-arz test set, results of our investigation show that AIDA deteriorated performance on SOOV happens due to the noise that its MSA replacements add to the non-SOOV proportion of data. To assess this hypothesis we ran the best experimental setup (decoding:lex-to-lex:lem+POS-to-lex, training: MSA+DA) on the proportion of sentences in BOLT-arz which contain at least one SOOV

Replacement Scheme	DA COOV		SOOV	setup	Enhanced Sentences DA COOV	SOOV
AIDA Replacement	type	21	712	lex-to-lex	40%	58%
				lem+POS-to-lex	**60%**	35%
	token	26	1481	lex-to-lex:lem+POS-to-lex	55%	57%
MADAMIRA Replacement	type	9	194	lex-to-lex	34%	55%
				lem+POS-to-lex	34%	47%
	token	9	281	lex-to-lex:lem+POS-to-lex	45%	**62%**

Table 5: Columns from left to right: number of DA COOV, SOOV and percentages of enhanced sentences for BOLT-arz set.

Replacement Scheme	DA COOV		SOOV	setup	Enhanced Sentences DA COOV	SOOV
AIDA Replacement	type	6	376	lex-to-lex	67%	44%
				lem+POS-to-lex	**84%**	35%
	token	6	499	lex-to-lex:lem+POS-to-lex	50%	**61%**
MADAMIRA Replacement	type	7	559	lex-to-lex	29%	40%
				lem+POS-to-lex	27%	34%
	token	7	852	lex-to-lex:lem+POS-to-lex	43%	48%

Table 6: Similar to Table 5 for MT09 set.

word as processed using AIDA and MADAMIRA (the intersection subset). It is worth noting that compared to the baseline BLEU score of 23.8 on this subset, AIDA achieves a BLEU score of 24.4 while MADAMIRA only achieves a lower BLEU score of 24.0. This implicitly demonstrates that AIDA provides better MSA equivalents even for DA words which have MSA homographs with different meanings (faux amis cases). Overall, we note that the same results can be captured from Table 2 that shows AIDA outperforming MADAMIRA in identifying and replacing DA words.

6 Conclusion and Future Work

We presented a new approach to enhance DA-to-EN machine translation by reducing the rate of DA OOV words. We employed AIDA and MADAMIRA to identify DA words and replace them with the corresponding MSA equivalent.
We showed our replacement scheme reaches a noticeable enhancement in SMT performance for SOOVs. This can be considered one of the contributions of this work which was not addressed in the previous studies before. The results of evaluation on two blind test sets showed that using AIDA to identify and replace MSA equivalents enhances translation results by 0.4% absolute BLEU (1.6% relative BLEU) and using MADAMIRA achieves 0.3% absolute BLEU (1.2% relative BLEU) enhancement over the baseline on two blind test sets. One of the interesting ideas to extend this project in the future is to combine AIDA and MADAMIRA top choices in a confusion network and feeding this confusion network to the SMT system. Acquiring more DA-to-EN parallel data to enrich translation models is another work which we intend to pursue later. Moreover, evaluating possible effects of different genres and domains on the framework efficiency provides another path to extend this work in future.

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

Abhaya Agarwal, and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-

correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation,* pp. 115-118,

Rania Al-Sabbagh and Roxana Girju. 2010. Mining the Web for the Induction of a dialectal Arabic Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC),*

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of EACL 2006,*

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Pradeep Dasigi, Heba Elfardy, Ramy Eskander, Nizar Habash, Abdelati Hawwari and Wael Salloum. 2014. Tharwa: A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon. In *Proceedings of LREC 2014,* Reykjavik, Iceland.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy and Yassin Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Proceedings of LREC Workshop on Semitic Language Processing,* pp. 6674.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in Arabic. In *Proceedings of ACL 2013,* Sofia, Bulgaria.

Heba Elfardy, Mohamed Al-Badrashiny and Mona Diab. 2013. Code Switch Point Detection in Arabic. In Natural Language Processing and Information Systems, Springer Berlin Heidelberg, pp. 412-416.

David Graff and Christopher Cieri. 2003. English Gigaword, LDC Catalog No.: LDC2003T05 Linguistic Data Consortium, University of Pennsylvania.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of ACL 2005,*

Nizar Habash. 2009. REMOOV: A tool for online handling of out-of-vocabulary words in machine translation.. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR),* Cairo, Egypt.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of ACL 2008: HLT, Short Papers,* Columbus, Ohio.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In *Proceedings of NAACL 2013:HLT,* pp. 426-432.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL 2007, Demo and Poster Sessions.* Prague, Czech Republic.

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,* Cairo, Egypt.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of ACL 2003,* pages 160-167 Sapporo, Japan.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational linguistics. Vol. 29. pp. 19-51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002,* pages 311318, Philadelphia, PA.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of LREC 2014,* Reykjavik, Iceland.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of NAACL 2013:HLT,* Atlanta, Georgia.

Wael Salloum and Nizar Habash. 2012. Elissa: A Dialectal to Standard Arabic Machine Translation System. In *Proceedings of COLING 2012,* Denver, Colorado.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of AMTA 2010,* Denver, Colorado.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing,* pp. 44-49.

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 AMTA 2006,* pp. 223-231.

Andreas Stolcke. 2002. SRILM an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing,*

Ying Zhang and Stephan Vogel. 2010. Significance Tests of Automatic Machine Translation Evaluation Metrics. In Machine Translation, Vol. 24, Issue 1, pages 51-65.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of NAACL 2012:HLT,*