Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic

Wael Salloum and **Nizar Habash**
Center for Computational Learning Systems
Columbia University
{wael,habash}@ccls.columbia.edu

Abstract

Modern Standard Arabic (MSA) has a wealth of natural language processing (NLP) tools and resources. In comparison, resources for dialectal Arabic (DA), the unstandardized spoken varieties of Arabic, are still lacking. We present ELISSA, a machine translation (MT) system for DA to MSA. ELISSA employs a rule-based approach that relies on morphological analysis, transfer rules and dictionaries in addition to language models to produce MSA paraphrases of DA sentences. ELISSA can be employed as a general preprocessor for DA when using MSA NLP tools. A manual error analysis of ELISSA's output shows that it produces correct MSA translations over 93% of the time. Using ELISSA to produce MSA versions of DA sentences as part of an MSA-pivoting DA-to-English MT solution, improves BLEU scores on multiple blind test sets between 0.6% and 1.4%.

1 Introduction

Much work has been done on Modern Standard Arabic (MSA) natural language processing (NLP) and machine translation (MT), especially Statistical MT (SMT). MSA has a wealth of resources in terms of morphological analyzers, disambiguation systems, and parallel corpora. In comparison, research on dialectal Arabic (DA), the unstandardized spoken varieties of Arabic, is still lacking in NLP in general and MT in particular. In this paper we present ELISSA, our DA-to-MSA MT system, and show how it can help improve the translation of highly dialectal Arabic text into English by pivoting on MSA.

The ELISSA approach can be summarized as follows. First, ELISSA uses different techniques to identify dialectal words and multi-word constructions (phrases) in a source sentence. Then, ELISSA produces MSA paraphrases for the selected words and phrase using a rule-based component that depends on the existence of a dialectal morphological analyzer, a list of morphosyntactic transfer rules, and DA-MSA dictionaries. The resulting MSA is in a lattice form that we pass to a language model for n-best decoding. The output of ELISSA, whether a top-1 choice sentence or n-best sentences, is passed to an MSA-English SMT system to produce the English translation sentence. ELISSA-based MSA-pivoting for DA-to-English SMT improves BLEU scores (Papineni et al., 2002) on three blind test sets between 0.6% and 1.4%. A manual error analysis of translated words shows that ELISSA produces correct MSA translations over 93% of the time.

The rest of this paper is structured as follows: Section 2 motivates the use of ELISSA to improve DA-English SMT with an example. Section 3 discusses some of the challenges associated with processing Arabic and its dialects. Section 4 presents related work. Section 5 details ELISSA and its approach and Section 6 presents results evaluating ELISSA under a variety of conditions.

2 Motivating Example

Table 1 shows a motivating example of how pivoting on MSA can dramatically improve the translation quality of a statistical MT system that is trained on mostly MSA-to-English parallel corpora. In this example, we use Google Translate's online Arabic-English SMT system.[1] The table is divided into two parts. The top part shows a dialectal (Levantine) sentence, its reference translation to English, and its Google Translate translation. The Google Translate translation clearly struggles with most of the DA words, which were probably unseen in the training data (i.e., out-of-vocabulary – OOV) and were con-

[1] The system was used on February 21, 2013.

Proceedings of NAACL-HLT 2013, pages 348–358,
Atlanta, Georgia, 9–14 June 2013.

DA source	بهالحالة هاي ما حيكتبولو عحيط الصفحه الشخصية تبعو ولا بدن ياه يبعتلن كومينتات لأنو ماخبرهون ايمتا رح يروح عالبلد. *bhAlHAlħ hAy mA Hyktbwlw ςHyT AlSfHh AlšxSyħ tbςw wlA bdn yAh ybςtln kwmyntAt lÂnw mAxbrhwn AymtA rH yrwH ςAlbld.*
Human Reference	In this case, they will not write on his profile wall and they do not want him to send them comments because he did not tell them when he will go to the country.
Google Translate	Bhalhalh Hi Hictpoulo Ahat Profile Tbau not hull Weah Abatln Comintat Anu Mabarhun Oamta welcomed calls them Aalbuld.
Human DA-to-MSA	في هذه الحالة لن يكتبوا له على حائط صفحته الشخصية ولا يريدونه أن يرسل لهم تعليقات لأنه لم يخبرهم متى سيذهب إلى البلد. *fy hðh AlHAlħ ln yktbwA lh ςlý HAŷT SfHth AlšxSyħ wlA yrydwnh Ân yrsl lhm tςlyqAt lÂnh lm yxbrhm mtý syðhb Ălý Albld.*
Google Translate	In this case it would not write to him on the wall of his own and do not want to send their comments because he did not tell them when going to the country.

Table 1: A motivating example for DA-to-English MT by pivoting (bridging) on MSA. The top half of the table displays a DA sentence, its human reference translation and the output of Google Translate. The bottom half of the table shows the result of human translation into MSA of the DA sentence before sending it to Google Translate.

sidered proper nouns (transliterated and capitalized). The lack of DA-English parallel corpora suggests pivoting on MSA can improve the translation quality. In the bottom part of the table, we show a human MSA translation of the DA sentence above and its Google translation. We see that the results are quite promising. The goal of ELISSA is to model this DA-MSA translation automatically. In Section 5.4, we revisit this example to discuss ELISSA's performance on it. We show its output and its corresponding Google translation in Table 3.

3 Challenges for Processing Arabic and its Dialects

Contemporary Arabic is in fact a collection of varieties: MSA, the official language of the Arab World, which has a standard orthography and is used in formal settings; and DAs, the commonly used informal native varieties, which have no standard orthographies but have an increasing presence on the web. Arabic, in general, is a morphologically complex language which has rich inflectional morphology, expressed both templatically and affixationally, and several classes of attachable clitics. For example, the Arabic word وسيكتبونها *w+s+y-ktb-wn+hA*[2] 'and they will write it' has two proclitics (+و *w+* 'and' and +س *s+* 'will'), one prefix -ي *y-* '3rd person', one suffix -ون *-wn* 'masculine plural' and one pronominal enclitic +ها *+hA* 'it/her'. DAs differ from MSA phonologically, morphologically and to a lesser degree syntactically. The morphological differences are most noticeably expressed in the use of clitics and affixes that do not exist in MSA. For instance, the Levantine Arabic equivalent of the MSA example above is وحيكتبوها *w+H+y-ktb-w+hA* 'and they will write it'. The optionality of vocalic diacritics helps hide some of the differences resulting from vowel changes; compare the diacritized forms: Levantine *wHayikitbuwhA* and MSA *wasayaktubuwnahA*.

All of the NLP challenges of MSA (e.g., optional diacritics and spelling inconsistency) are shared by DA. However, the lack of standard orthographies for the dialects and their numerous varieties pose new challenges. Additionally, DAs are rather impoverished in terms of available tools and resources compared to MSA, e.g., there is very little parallel DA-English corpora and almost no MSA-DA parallel corpora. The number and sophistication of morphological analysis and disambiguation tools in DA is very limited in comparison to MSA (Duh and Kirchhoff, 2005; Habash and Rambow, 2006; Abo Bakr et al., 2008; Habash, 2010; Salloum and Habash, 2011; Habash et al., 2012; Habash et al., 2013). MSA tools cannot be effectively used to handle DA, e.g., Habash and Rambow (2006) report that over one-third of Levantine verbs cannot be analyzed using an MSA morphological analyzer.

[2] Arabic transliteration throughout the paper is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order) *AbtθjHxdðrzsšSDTĎςγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ă إ, Ā آ, ŵ ؤ, ŷ ئ, ħ ة, ý ى.

4 Related Work

Dialectal Arabic NLP. Several researchers have explored the idea of exploiting existing MSA rich resources to build tools for DA NLP (Chiang et al., 2006). Such approaches typically expect the presence of tools/resources to relate DA words to their MSA variants or translations. Given that DA and MSA do not have much in terms of parallel corpora, rule-based methods to translate DA-to-MSA or other methods to collect word-pair lists have been explored. For example, Abo Bakr et al. (2008) introduced a hybrid approach to transfer a sentence from Egyptian Arabic into MSA. This hybrid system consisted of a statistical system for tokenizing and tagging, and a rule-based system for constructing diacritized MSA sentences. Moreover, Al-Sabbagh and Girju (2010) described an approach of mining the web to build a DA-to-MSA lexicon. In the context of DA-to-English SMT, Riesa and Yarowsky (2006) presented a supervised algorithm for online morpheme segmentation on DA that cut the OOV words by half.

Machine Translation for Closely Related Languages. Using closely related languages has been shown to improve MT quality when resources are limited. Hajič et al. (2000) argued that for very close languages, e.g., Czech and Slovak, it is possible to obtain a better translation quality by using simple methods such as morphological disambiguation, transfer-based MT and word-for-word MT. Zhang (1998) introduced a Cantonese-Mandarin MT that uses transformational grammar rules. In the context of Arabic dialect translation, Sawaf (2010) built a hybrid MT system that uses both statistical and rule-based approaches for DA-to-English MT. In his approach, DA is normalized into MSA using a dialectal morphological analyzer. In previous work, we presented a rule-based DA-MSA system to improve DA-to-English MT (Salloum and Habash, 2011; Salloum and Habash, 2012). Our approach used a DA morphological analyzer (ADAM) and a list of hand-written morphosyntactic transfer rules. This use of "resource-rich" related languages is a specific variant of the more general approach of using pivot/bridge languages (Utiyama and Isahara, 2007; Kumar et al., 2007). In the case of MSA and DA variants, it is plausible to consider the MSA variants of a DA phrase as monolingual paraphrases (Callison-Burch et al., 2006; Du et al., 2010). Also related is the work by Nakov and Ng (2011), who use morphological knowledge to generate paraphrases for a morphologically rich language, Malay, to extend the phrase table in a Malay-to-English SMT system.

Pivoting on MSA or acquiring more DA-English data? Zbib et al. (2012) demonstrated an approach to cheaply obtaining DA-English data. They used Amazon's Mechanical Turk (MTurk) to create a DA-English parallel corpus of 1.5M words and added it to a 150M MSA-English parallel corpus to create the training corpus of their SMT system. They also used MTurk to translate their dialectal test set to MSA in order to compare the MSA-pivoting approach to the direct translation from DA to English approach. They showed that even though pivoting on MSA (produced by Human translators in an oracle experiment) can reduce OOV rate to 0.98% from 2.27% for direct translation (without pivoting), it improves by 4.91% BLEU while direct translation improves by 6.81% BLEU over their 12.29% BLEU baseline (direct translation using the 150M MSA system). They concluded that simple vocabulary coverage is not sufficient and the domain mismatch is a more important problem. The approach we take in this paper is orthogonal to such efforts to build parallel data. We plan to study interactions between the two types of solutions in the future.

Our work is most similar to Sawaf (2010)'s MSA-pivoting approach. In his approach, DA is normalized into MSA using character-based DA normalization rules, a DA morphological analyzer, a DA normalization decoder that relies on language models, and a lexicon. Similarly, we use some character normalization rules, a DA morphological analyzer, and DA-MSA dictionaries. In contrast, we use hand-written morphosyntactic transfer rules that focus on translating DA morphemes and lemmas to their MSA equivalents.

In our previous work (Salloum and Habash, 2011; Salloum and Habash, 2012), we applied our approach to tokenized Arabic and our DA-MSA transfer component used feature transfer rules only. We did not use a language model to pick the best path; instead we kept the ambiguity in the lattice and passed it to our SMT system. In contrast, in this paper, we run ELISSA on untokenized Arabic, we use

feature, lemma, and surface form transfer rules, and we pick the best path of the generated MSA lattice through a language model.

Certain aspects of our approach are similar to Riesa and Yarowsky (2006)'s, in that we use morphological analysis for DA to help DA-English MT; but unlike them, we use a rule-based approach to model DA morphology.

5 ELISSA

ELISSA is a DA-to-MSA MT System. ELISSA uses a rule-based approach (with some statistical components) that relies on the existence of a DA morphological analyzer, a list of hand-written transfer rules, and DA-MSA dictionaries to create a mapping of DA to MSA words and construct a lattice of possible sentences. ELISSA uses a language model to rank and select the generated sentences.

ELISSA supports untokenized (raw) input only. ELISSA supports three types of output: top-1 choice, an n-best list or a map file that maps source words/phrases to target phrases. The top-1 and n-best lists are determined using an untokenized MSA language model to rank the paths in the MSA translation output lattice. This variety of output types makes it easy to plug ELISSA with other systems and to use it as a DA preprocessing tool for other MSA systems, e.g., MADA (Habash and Rambow, 2005) or AMIRA (Diab et al., 2007).

ELISSA's approach consists of three major steps preceded by a *preprocessing and normalization* step, that prepares the input text to be handled (e.g., UTF-8 cleaning, Alif/Ya normalization, word-lengthening normalization), and followed by a *post-processing* step, that produces the output in the desired form (e.g., encoding choice). The three major steps are **Selection**, **Translation**, and **Language Modeling**.

5.1 Selection

In the first step, ELISSA identifies which words or phrases to paraphrase and which words or phrases to leave as is. ELISSA provides different methods (techniques) for selection, and can be configured to use different subsets of them. In Section 6 we use the term "selection mode" to denote a subset of selection methods. Selection methods are classified into *Word-based selection* and *Phrase-based selection*.

Word-based selection. Methods of this type fall in the following categories:

a. User token-based selection: The user can mark specific words for selection using the tag '/DIA' (stands for 'dialect') after each word to select.

b. User type-based selection: The user can specify a list of words to select from, e.g., OOVs. Also the user can provide a list of words and their frequencies and specify a cut-off threshold to prevent selecting a frequent word.

c. Morphology-based word selection: ELISSA uses ADAM (Salloum and Habash, 2011) to select words that have DA analyses only (DIAONLY) or DA/MSA analyses (DIAMSA).

d. Dictionary-based selection: ELISSA selects words based on their existence in the DA side of our DA-MSA dictionaries.

e. All: ELISSA selects every word in an input sentence.

Phrase-based selection. This selection type uses hand-written rules to identify dialectal multi-word constructions that are mappable to single or multi-word MSA constructions. The current count of these rules is 25. Table 2 presents some rule categories and related examples.

In the current version of ELISSA, words can be selected using either the phrase-based selection method or a word-based selection method, but not both. Phrase-based selection has precedence. We evaluate different settings for selection step in Section 6.

5.2 Translation

In this step, ELISSA translates the selected words and phrases to their MSA equivalent paraphrases. The specific type of selection determines the type of the translation, e.g., phrase-based selected words are translated using phrase-based translation rules. The MSA paraphrases are then used to form an MSA lattice.

Word-based translation. This category has two types of translation techniques: *surface translation* that uses DA-to-MSA surface-to-surface (S2S) transfer rules (TRs) and *deep (morphological) translation* that uses the classic rule-based machine translation flow: analysis, transfer and generation. The

Rule Category	Selection Examples	Translation Examples
Dialectal Idafa	الجيش الوطني بتاعنا *Aljyš AlwTny btAçnA* 'the-army the-national ours'	جيشنا الوطني *jyšnA AlwTny* 'our-army the-national'
Verb + *flipped direct and indirect objects*	حضرلها ياهن *HDrlhA yAhn* 'he-prepared-for-her them'	حضرهم لها *HDrhm lhA* 'he-prepared-them for-her'
Special dialectal expressions	بدو اياها *bdw AyAhA* 'his-desire her'	يريدها *yrydhA* 'he-desires-her'
Negation + verb	وما حيكتبولو *wmA Hyktbwlw* 'and-not they-will-write-to-him'	ولن يكتبوا له *wln yktbwA lh* 'and-will-not they-write to-him'
Negation + agent noun	فمش لاقية *fmš lAqyħ* 'so-not finding'	فلا تجد *flA tjd* 'so-not she-finds'
Negation + closed-class words	ما عدكم *mA çdkm* 'not with-you'	ليس لديكم *lys ldykm* 'not with-you'

Table 2: Examples of some types of phrase-based selection and translation rules.

DA Phrase	وما راحولا *wmA rAHwlA* 'And they did not go to her'				
Analysis	Word 1		Word 2		
	Proclitics	[Lemma & Features]	[Lemma & Features]	[Lemma & Features]	Enclitic
	w+ conj+ and+	mA [neg] not	rAHw [rAH PV subj:3MP] they go	+l +prep +to	+A +pron$_{3FS}$ +her
Transfer	Word 1		Word 2	Word 3	
	Proclitics	[Lemma & Features]	[Lemma & Features]	[Lemma & Features]	Enclitic
	conj+ and+	[lam] did not	[ðahab IV subj:3MP] they go	[Ălý] to	+pron$_{3FS}$ +her
Generation	w+	lm	yðhbwA	Ăly	+hA
MSA Phrase	ولم يذهبوا إليها *wlm yðhbwA ĂlyhA* 'And they did not go to her'				

Figure 1: An example illustrating the analysis-transfer-generation steps to translate a dialectal multi-word expression into its MSA equivalent phrase.

dialectal morphological analysis step uses ADAM (Salloum and Habash, 2011) to get a list of dialectal analyses. The morphosyntactic transfer step uses lemma-to-lemma (L2L) and features-to-features (F2F) transfer rules to change lemmas, clitics or features, and even split up the dialectal word into multiple MSA word analyses (such as splitting negation words and indirect objects). The MSA morphological generation step uses the general tokenizer/generator TOKAN (Habash, 2007) to generate untokenized surface form words. For more details, see Salloum and Habash (2011).

Phrase-based translation. Unlike the word-based translation techniques which map single DA words to single or multi-word MSA sequences, this technique uses hand-written multi-word transfer rules that map multi-word DA constructions to single or multi-word MSA constructions. In the current system, there are 47 phrase-based transfer rules. Many of the word-based morphosyntactic transfer rules are re-used for phrase-based translation. Figure 1 shows an example of a phrase-based morphological translation of the two-word DA sequence وما راحولا *wmA rAHwlA* 'And they did not go to her'. If these two words were spelled as a single word, ومارحولا *wmArAHwlA*, we would still get the same result using the word-based translation technique only. Table 2 shows some rule categories along with selection and translation examples.

5.3 Language Modeling

The language model (LM) component uses the SRILM lattice-tool for weight assignment and n-best decoding (Stolcke, 2002). ELISSA comes with a default 5-gram LM file trained on ∼200M unto-

DA source	(بهالحالة هاي) [1] (ما حيكتبولو) [2] عحيط [3] (الصفحه الشخصية تبعو) [4] ولا (بدن ياه) [5] يبعتلن [6] كومينتات [7] لأنو [8] ماخبرهون [9] ايمتا [10] (رح يروح) [11] عالبلد [12]. *(bhAlHAlħ hAy)[1] (mA Hyktbwlw)[2] ʕHyT[3] (AlSfHh AlšxSyħ tbʕw)[4] wlA (bdn yAh)[5] ybʕtln[6] kwmyntAt[7] lÂnw[8] mAxbrhwn[9] AymtA[10] (rH yrwH)[11] ʕAlbld[12].*
Human Reference	In this case, they will not write on his profile wall and they do not want him to send them comments because he did not tell them when he will go to the country.
Google Translate	Bhalhalh Hi Hictpoulo Ahat Profile Tbau not hull Weah Abatln Comintat Anu Mabarhun Oamta welcomed calls them Aalbuld.
ELISSA **DA-to-MSA**	(في هذه الحالة) [1] (لن يكتبوا له) [2] (علي حائط) [3] (صفحته الشخصية) [4] ولا (يريدونه ان) [5] (يرسل اليهم) [6] تعليقات [7] لانه [8] (لم يخبرهم) [9] متي [10] سيذهب [11] (الي البلد) [12] . *(fy hðh AlHAlħ)[1] (ln yktbwA lh)[2] (ʕly HAŷT)[3] (SfHth AlšxSyħ)[4] wlA (yrydwnh An)[5] (yrsl Alyhm)[6] tʕlyqAt[7] lAnh[8] (lm yxbrhm)[9] mty[10] syðhb[11] (Aly Albld)[12].*
Google Translate	In this case it would not write to him on the wall of his own and do not want to send them comments that he did not tell them when going to the country.

Table 3: Revisiting our motivating example, but with ELISSA-based DA-to-MSA middle step. ELISSA's output is Alif/Ya normalized. Parentheses are added for illustrative reasons to highlight how multi-word DA constructions are selected and translated. Superscript indices link the selected words and phrases with their MSA translations.

kenized Arabic words of Arabic Gigaword (Parker et al., 2009). Users can specify their own LM file and/or interpolate it with our default LM. This is useful for adapting ELISSA's output to the Arabic side of the training data.

5.4 Revisiting our Motivating Example

We revisit our motivating example in Section 2 and show automatic MSA-pivoting through ELISSA. Table 3 is divided into two parts. The first part is copied from Table 1 for convenience. The second part shows ELISSA's output on the dialectal sentence and its Google Translate translation. The produced MSA is not perfect, but is clearly an improvement over doing nothing as far as usability for MT into English.

6 Evaluation

In this section, we present two evaluations of ELISSA. The first is an extrinsic evaluation of ELISSA as part of MSA-pivoting for DA-to-English SMT. And the second is an intrinsic evaluation of the quality of ELISSA's MSA output.

6.1 DA-English MT Evaluation

6.1.1 Experimental Setup

We use the open-source Moses toolkit (Koehn et al., 2007) to build a phrase-based SMT system trained on mostly MSA data (64M words on the Arabic side) obtained from several LDC corpora including some limited DA data. Our system uses a standard phrase-based architecture. The parallel corpus is word-aligned using GIZA++ (Och and Ney, 2003). Phrase translations of up to 10 words are extracted in the Moses phrase table. The language model for our system is trained on the English side of the bitext augmented with English Gigaword (Graff and Cieri, 2003). We use a 5-gram language model with modified Kneser-Ney smoothing. Feature weights are tuned to maximize BLEU on the NIST MTEval 2006 test set using Minimum Error Rate Training (Och, 2003). This is only done on the baseline systems. The English data is tokenized using simple punctuation-based rules. The Arabic side is segmented according to the Arabic Treebank (ATB) tokenization scheme (Maamouri et al., 2004) using the MADA+TOKAN morphological analyzer and tokenizer v3.1 (Habash and Rambow, 2005; Roth et al., 2008). The Arabic text is also Alif/Ya normalized. MADA-produced Arabic lemmas are used for word alignment.

We use the same development (dev) and test sets used by Salloum and Habash (2011) (we will call them speech-dev and speech-test, respectively) and we compare to them in the next sections. We also evaluate on two web-crawled blind test sets: the Levantine test set presented in Zbib et al. (2012) (we will call it web-lev-test) and the Egyptian Dev-MT-v2 development data of the DARPA BOLT program (we will call it web-egy-test). The speech-dev set has 1,496 sentences with 32,047 untokenized Arabic words. The speech-test set has 1,568 sentences with

32,492 untokenized Arabic words. The web-levtest set has 2,728 sentences with 21,179 untokenized Arabic words. The web-egy-test set has 1,553 sentences with 21,495 untokenized Arabic words. The two speech test sets contain multi-dialect (e.g., Iraqi, Levantine, Gulf, and Egyptian) broadcast conversational (BC) segments (with three reference translations), and broadcast news (BN) segments (with only one reference, replicated three times). The web-egy-test has two references while the web-levtest has only one reference. Results are presented in terms of BLEU (Papineni et al., 2002). All evaluation results are case insensitive.

6.1.2 Results on the Development Set

We experimented with different method combinations in the selection and translation components in ELISSA. We use the term selection mode and translation mode to denote a certain combination of methods in selection or translation, respectively. Due to limited space, we only present the best selection mode variation experiments. Other selection modes were tried but they proved to be consistently lower than the rest. The 'F2F+L2L; S2S' word-based translation mode (using morphological transfer of features and lemmas along with surface form transfer) showed to be consistently better than other method combinations across all selection modes. In this paper we only use 'F2F+L2L; S2S' word-based translation mode. Phrase-based translation mode is used when phrase-based selection mode is used.

To rank paraphrases in the generated MSA lattice, we combine two 5-gram untokenized Arabic language models: one is trained on Arabic Gigaword data and the other is trained the Arabic side of our SMT training data. The use of the latter LM gave frequent dialectal phrases a higher chance to appear in ELISSA's output; thus, making the output "more dialectal" but adapting it to our SMT input. Experiments showed that using both LMs is better than using each one alone.

In all the experiments, we run the DA sentence through ELISSA to generate a top-1 MSA translation, which we then tokenize through MADA before sending to the MSA-English SMT system. Our baseline is to not run ELISSA at all; instead, we send the DA sentence through MADA before applying the MSA-English MT system.

Table 4 summarizes the experiments and results on the dev set. The rows of the table are the different systems (baseline and ELISSA's experiments). All differences in BLEU scores from the baseline are statistically significant above the 95% level. Statistical significance is computed using paired bootstrap re-sampling (Koehn, 2004). The name of the system in ELISSA's experiments denotes the combination of selection method. ELISSA's experiments are grouped into three groups: simple selection, frequency-based selection, and phrase-based selection. Simple selection group consists of five systems: OOV, ADAM, OOV U ADAM, DICT, and OOV U ADAM U DICT. The OOV selection mode identifies the untokenized OOV words. In the ADAM selection mode, or the morphological selection mode, we use ADAM to identify dialectal words. Experiments showed that ADAM's DIAMSA mode (selecting words that have at least one dialectal analysis) is slightly better than ADAM's DIAONLY mode (selecting words that have only dialectal analyses and no MSA ones). The OOV U ADAM selection mode is the union of the OOVs and ADAM selection modes. In DICT selection mode, we select dialectal words that exist in our DA-MSA dictionaries. The OOV U ADAM U DICT selection mode is the union of the OOVs, ADAM, and DICT selection modes. The results show that combining the output of OOV selection method and ADAM selection method is the best. DICT selection method hurts the performance of the system when used because dictionaries usually have frequent dialectal words that the SMT system already knows how to handle.

In the frequency-based selection group, we exclude from word selection all words with number of occurrences in the training data that is above a certain threshold. This threshold was determined empirically to be 50. The string '- (Freq >= 50)' means that all words with frequencies of 50 or more should not be selected. The results show that excluding frequent dialectal words improves the best simple selection system. It also shows that using DICT selection improves the best system if frequent words are excluded.

In the last system group, phrase+word-based selection, phrase-based selection is used to select phrases and add them on top of the best performers of the previous two groups. Phrase-based trans-

Test Set	speech-dev	
	BLEU	Diff.
Baseline	37.20	0.00
Select: OOV	37.75	0.55
Select: ADAM	37.88	0.68
Select: OOV U ADAM	37.89	0.69
Select: DICT	37.06	-0.14
Select: OOV U ADAM U DICT	37.53	0.33
Select: (OOV U ADAM) - (Freq >= 50)	37.96	0.76
Select: (OOV U ADAM U DICT) - (Freq >= 50)	38.00	0.80
Select: Phrase; (OOV U ADAM)	37.99	0.79
Select: Phrase; ((OOV U ADAM) - (Freq >= 50))	38.05	0.85
Select: Phrase; ((OOV U ADAM U DICT) - (Freq >= 50))	**38.10**	**0.90**

Table 4: Results for the speech-dev set in terms of BLEU. The 'Diff.' column shows result differences from the baseline. The rows of the table are the different systems (baseline and ELISSA's experiments). The name of the system in ELISSA's experiments denotes the combination of selection method. In all ELISSA's experiments, all word-based translation methods are tried. Phrase-based translation methods are used when phrase-based selection is used (i.e., the last three rows). The best system is in bold.

lation is also added to word-based translation. Results show that selecting and translating phrases improve the three best performers of word-based selection. The best performer, shown in the last raw, suggests using phrase-based selection and restricted word-based selection. The restriction is to include OOV words and selected low frequency words that have at least one dialectal analysis or appear in our dialectal dictionaries. Comparing the best performer to the OOV selection mode system shows that translating low frequency in-vocabulary dialectal words and phrases to their MSA paraphrases can improve the English translation. This is a similar conclusion to our previous work in Salloum and Habash (2011).

6.1.3 Results on the Blind Test Sets

We run the system settings that performed best on the dev set along with the OOV selection mode system on the three blind test set. Results and their differences from the baseline are reported in Table 5. We see that OOV selection mode system always improves over the baseline for all test sets. Also, the best performer on the dev is the best performer for all test sets. The improvements of the best performer over the OOV selection mode system on all test sets confirm that translating low frequency in-vocabulary dialectal words and phrases to their MSA paraphrases can improve the English translation. Its improvements over the baseline for the three test sets are: 0.95% absolute BLEU (or 2.5% relative) for the speech-test, 1.41% absolute BLEU (or 15.4% relative) for the web-lev-test, and 0.61% absolute BLEU (or 3.2% relative) for the web-egy-test.

6.1.4 A Case Study

We next examine an example in some detail. Table 6 shows a dialectal sentence along with its ELISSA's translation, English references, the output of the baseline system and the output of our best system. The example shows a dialectal word هالمبلغ *hAlmblγ* 'this-amount/sum', which is not translated by the baseline (although it appears in the training data, but quite infrequently such that all of its phrase table occurrences have restricted contexts, making it effectively an OOV). The dialectal proclitic هال+ *hAl+* 'this-' comes sometimes in the dialectal construction: 'hAl+NOUN DEM' (as in this example: هالمبلغ هذا *hAlmblγ hðA* 'this-amount/sum this'). ELISSA's selection component captures this multi-word expression and its translation component produces the following paraphrases: هذا المبلغ *hðA Almblγ* 'this amount/sum' (*hðA* is used with masculine singular nouns), هذه المبلغ *hðh Almblγ* 'this amount/sum' (*hðh* is used with feminine singular or irrational plural nouns), and هؤلاء المبلغ *hŵlA' Almblγ* 'these amount/sum' (*hŵlA'* is used with rational plural nouns). ELISSA's language modeling component picks the first MSA paraphrase, which perfectly fits the context and satisfies the gender/number/rationality agreement (note that the word *Almblγ* is an irrational masculine singular

Test Set	speech-test		web-lev-test		web-egy-test	
	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.
Baseline	38.18	0.00	9.13	0.00	18.98	0.00
Select: OOV	38.76	0.58	9.65	0.62	19.19	0.21
Select: Phrase; ((OOV U ADAM U DICT) - (Freq >= 50))	**39.13**	**0.95**	**10.54**	**1.41**	**19.59**	**0.61**

Table 5: Results for the three blind test sets (table columns) in terms of BLEU. The 'Diff.' columns show result differences from the baselines. The rows of the table are the different systems (baselines and ELISSA's experiments). The best systems are in bold.

noun). For more on Arabic morpho-syntactic agreement patterns, see Alkuhlani and Habash (2011). Finally, the best system translation for the selected phrase is 'this sum'. We can see how both the accuracy and fluency of the sentence have improved.

DA sentence	fmA mA AtSwr **hAlmblγ hðA** yςny.
ELISSA's output	fmA mA AtSwr **hðA Almblγ** yςny.
References	I don't think **this amount** is I mean. So I do not I do not think **this cost** I mean. So I do not imagine **this sum** I mean
Baseline	So i don't think **hAlmblg this** means.
Best system	So i don't think **this sum** i mean.

Table 6: An example of handling dialectal words/phrases using ELISSA and its effect on the accuracy and fluency of the English translation. Words of interest are bolded.

6.2 DA-to-MSA Translation Quality

We conducted a manual error analysis comparing ELISSA's input (the original dev set) to its output using our best system settings from the experiments above. Out of 708 affected sentences, we randomly selected 300 sentences (42%). Out of the 482 handled tokens, 449 (93.15%) tokens have good MSA translations, and 33 (6.85%) tokens have wrong MSA translations. Most of the wrong translations are due to spelling errors, proper nouns, and weak input sentence fluency (especially due to speech effect). This analysis clearly validates ELISSA's MSA output. Of course, a correct MSA output can still be mistranslated by the MT system we used above if it is not in the vocabulary of the MT system.

7 Conclusion and Future Work

We presented ELISSA, a tool for DA-MSA translation. ELISSA employs a rule-based MT approach that relies on morphological analysis, transfer rules and dictionaries in addition to language models to produce MSA paraphrases of dialectal sentences. Using ELISSA to produce MSA versions of dialectal sentences as part of an MSA-pivoting DA-to-English MT solution, improves BLEU scores on three blind test sets by: 0.95% absolute BLEU (or 2.5% relative) for a speech multi-dialect (Iraqi, Levantine, Gulf, Egyptian) test set, 1.41% absolute BLEU (or 15.4% relative) for a web-crawled Levantine test set, and 0.61% absolute BLEU (or 3.2% relative) for a web-crawled Egyptian test set. A manual error analysis of translated selected words shows that our system produces correct MSA translations over 93% of the time.

In the future, we plan to extend ELISSA's coverage of phenomena in the handled dialects and to new dialects. We also plan to automatically learn additional rules from limited available data (DA-MSA or DA-English). We also would like to do additional MT experiments where we use ELISSA to preprocess the training data, comparable to experiments done by Sawaf (2010). We are interested in studying how our approach can be combined with solutions that simply add more dialectal training data since the two directions are complementary in that they address linguistic normalization and domain coverage. Finally, we look forward to experimenting with ELISSA as a preprocessing system for a variety of dialect NLP applications similar to Chiang et al. (2006)'s work on dialect parsing, for example.

ELISSA will be publicly available. Please contact the authors for more information.

Acknowledgment

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-12-C-0014. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

References

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

Rania Al-Sabbagh and Roxana Girju. 2010. Mining the Web for the Induction of a Dialectical Arabic Lexicon. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

Sarah Alkuhlani and Nizar Habash. 2011. A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon, USA.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of the European Chapter of ACL (EACL)*.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2007. Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking. In Antal van den Bosch and Abdelhadi Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Kluwer/Springer.

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP'10, pages 420–429, Cambridge, Massachusetts.

Kevin Duh and Katrin Kirchhoff. 2005. POS tagging of dialectal Arabic: a minimally supervised approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Semitic '05, pages 55–62, Ann Arbor, Michigan.

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 the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Nizar Habash. 2007. Arabic Morphological Representations for Machine Translation. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Jan Hajič, Jan Hric, and Vladislav Kubon. 2000. Machine Translation of Very Close Languages. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'2000)*, pages 7–12, Seattle.

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 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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference*

on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 42–50, 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*, pages 102–109, Cairo, Egypt.

Preslav Nakov and Hwee Tou Ng. 2011. Translating from Morphologically Complex Languages: A Paraphrase-Based Approach. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'2011)*, Portland, Oregon, USA.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics*, pages 160–167, 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 the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2009. Arabic Gigaword Fourth Edition. LDC catalog number No. LDC2009T30, ISBN 1-58563-532-4.

Jason Riesa and David Yarowsky. 2006. Minimally Supervised Morphological Segmentation with Applications to Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA06)*, pages 185–192, Cambridge,MA.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of ACL-08: HLT, Short Papers*, pages 117–120, Columbus, Ohio.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Wael Salloum and Nizar Habash. 2012. Elissa: A Dialectal to Standard Arabic Machine Translation System. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Demonstration Papers*, pages 385–392, Mumbai, India.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado.

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

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491.

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 the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada, June. Association for Computational Linguistics.

Xiaoheng Zhang. 1998. Dialect MT: a case study between Cantonese and Mandarin. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 1460–1464, Montreal, Canada.