Machine Translation of Arabic Dialects

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas,
Richard Schwartz, John Makhoul, Omar F. Zaidan[†]**, Chris Callison-Burch**[‡]
Raytheon BBN Technologies, Cambridge MA
†Microsoft Research, Redmond WA
‡Johns Hopkins University, Baltimore MD

Abstract

Arabic Dialects present many challenges for machine translation, not least of which is the lack of data resources. We use crowdsourcing to cheaply and quickly build Levantine-English and Egyptian-English parallel corpora, consisting of 1.1M words and 380k words, respectively. The dialectal sentences are selected from a large corpus of Arabic web text, and translated using Amazon's Mechanical Turk. We use this data to build Dialectal Arabic MT systems, and find that small amounts of dialectal data have a dramatic impact on translation quality. When translating Egyptian and Levantine test sets, our Dialectal Arabic MT system performs 6.3 and 7.0 BLEU points higher than a Modern Standard Arabic MT system trained on a 150M-word Arabic-English parallel corpus.

1 Introduction

The Arabic language is a well-known example of *diglossia* (Ferguson, 1959), where the formal variety of the language, which is taught in schools and used in written communication and formal speech (religion, politics, etc.) differs significantly in its grammatical properties from the informal varieties that are acquired natively, which are used mostly for verbal communication. The spoken varieties of the Arabic language (which we refer to collectively as *Dialectal Arabic*) differ widely among themselves, depending on the geographic distribution and the socio-economic conditions of the speakers, and they diverge from the formal variety known as Modern Standard Arabic (MSA) (Embarki and Ennaji, 2011). Significant differences in the phonology, morphology, lexicon and even syntax render some of these varieties mutually incomprehensible.

The use of Dialectal Arabic has traditionally been confined to informal personal speech, while writing has been done almost exclusively using MSA (or its ancestor *Classical Arabic*). This situation is quickly changing, however, with the rapid proliferation of social media in the Arabic-speaking part of the world, where much of the communication is composed in dialect. The focus of the Arabic NLP research community, which has been mostly on MSA, is turning towards dealing with informal communication, with the introduction of the DARPA BOLT program. This new focus presents new challenges, the most obvious of which is the lack of dialectal linguistic resources. Dialectal text, which is usually user-generated, is also noisy, and the lack of standardized orthography means that users often improvise spelling. Dialectal data also includes a wider range of topics than formal data genres, such as newswire, due to its informal nature. These challenges require innovative solutions if NLP applications are to deal with Dialectal Arabic effectively.

In this paper:

- We describe a process for cheaply and quickly developing parallel corpora for Levantine-English and Egyptian-English using Amazon's Mechanical Turk crowdsourcing service (§3).
- We use the data to perform a variety of machine translation experiments showing the impact of morphological analysis, the limited value of adding MSA parallel data, the usefulness of cross-dialect training, and the effects of translating from dialect to MSA to English (§4).

We find that collecting dialect translations has a low cost ($0.03/word) and that relatively small amounts of data has a dramatic impact on translation quality. When trained on 1.5M words of dialectal data, our system performs 6.3 to 7.0 BLEU points higher than when it is trained on 100 times more MSA data from a mismatching domain.

2 Previous Work

Existing work on natural language processing of Dialectal Arabic text, including machine translation, is somewhat limited. Previous research on Dialectal Arabic MT has focused on normalizing dialectal input words into MSA equivalents before translating to English, and they deal with inputs that contain a limited fraction of dialectal words. Sawaf (2010) normalized the dialectal words in a hybrid (rule-based and statistical) MT system, by performing a combination of character- and morpheme-level mappings. They then translated the normalized source to English using a hybrid MT or alternatively a Statistical MT system. They tested their method on proprietary test sets, observing about 1 BLEU point (Papineni et al., 2002) increase on broadcast news/conversation and about 2 points on web text. Salloum and Habash (2011) reduced the proportion of dialectal out-of-vocabulary (OOV) words also by mapping their affixed morphemes to MSA equivalents (but did not perform lexical mapping on the word stems). They allowed for multiple morphological analyses, passing them on to the MT system in the form of a lattice. They tested on a subset of broadcast news and broadcast conversation data sets consisting of sentences that contain at least one region marked as non-MSA, with an initial OOV rate against an MSA training corpus of 1.51%. They obtained a 0.62 BLEU point gain. Abo Bakr et al. (2008) suggested another hybrid system to map Egyptian Arabic to MSA, using morphological analysis on the input and an Egyptian-MSA lexicon.

Other work that has focused on tasks besides MT includes that of Chiang et al. (2006), who built a parser for spoken Levantine Arabic (LA) transcripts using an MSA treebank. They used an LA-MSA lexicon in addition to morphological and syntactic rules to map the LA sentences to MSA. Riesa and Yarowsky (2006) built a statistical morphological segmenter for Iraqi and Levantine speech transcripts, and showed that they outperformed rule-based segmentation with small amounts of training. Some tools exist for preprocessing and tokenizing Arabic text with a focus on Dialectal Arabic. For example, MAGEAD (Habash and Rambow, 2006) is a morphological analyzer and generator that can analyze the surface form of MSA and dialect words into their root/pattern and affixed morphemes, or generate the surface form in the opposite direction.

Amazon's Mechanical Turk (MTurk) is becoming an essential tool for creating annotated resources for computational linguistics. Callison-Burch and Dredze (2010) provide an overview of various tasks for which MTurk has been used, and offer a set of best practices for ensuring high-quality data.

Zaidan and Callison-Burch (2011a) studied the quality of crowdsourced translations, by quantifying the quality of non-professional English translations of 2,000 Urdu sentences that were originally translated by the LDC. They demonstrated a variety of mechanisms that increase the translation quality of crowdsourced translations to near professional levels, with a total cost that is less than one tenth the cost of professional translation.

Zaidan and Callison-Burch (2011b) created the Arabic Online Commentary (AOC) dataset, a 52M-word monolingual dataset rich in dialectal content. Over 100k sentences from the AOC were annotated by native Arabic speakers on MTurk to identify the dialect level (and dialect itself) in each, and the collected labels were used to train automatic dialect identification systems. Although a large number of dialectal sentences were identified (41% of sentences), none were passed on to a translation phase.

3 Data Collection and Annotation

Following Zaidan and Callison-Burch (2011a,b), we use MTurk to identify Dialectal Arabic data and to create a parallel corpus by hiring non-professional translators to translate the sentences that were labeled as being dialectal. We had Turkers perform three steps for us: dialect classification, sentence segmentation, and translation.

Since Dialectal Arabic is much less common in written form than in spoken form, the first challenge is to simply find instances of written Dialectal Arabic. We draw from a large corpus of monolingual Arabic text (approximately 350M words) that was harvested from the web by the LDC, largely from weblog and online user groups.[1] Before presenting our data to annotators, we filter it to identify

[1] Corpora: LDC2006E32, LDC2006E77, LDC2006E90, LDC2007E04, LDC2007E44, LDC2007E102, LDC2008E41, LDC2008E54, LDC2009E14, LDC2009E93.

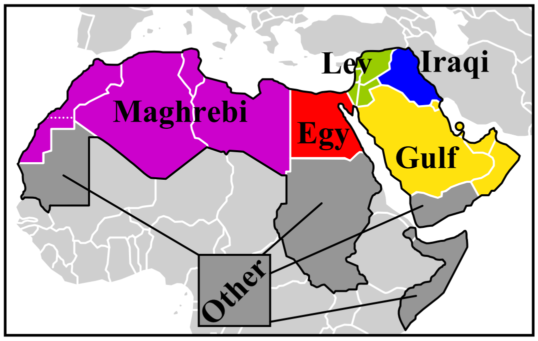

Figure 1: One possible breakdown of spoken Arabic into dialect groups: Maghrebi, Egyptian, Levantine, Gulf and Iraqi. Habash (2010) gives a breakdown along mostly the same lines. We used this map as an illustration for annotators in our dialect classification task (Section 3.1), with Arabic names for the dialects instead of English.

segments most likely to be dialectal (unlike Zaidan and Callison-Burch (2011b), who did no such prefiltering). We eliminate documents with a large percentage of non-Arabic or MSA words. We then retain documents that contain some number of dialectal words, using a set of manually selected dialectal words that was assembled by culling through the transcripts of the Levantine Fisher and Egyptian CallHome speech corpora. After filtering, the dataset contained around 4M words, which we used as a starting point for creating our Dialectal Arabic-English parallel corpus.

3.1 Dialect Classification

To refine the document set beyond our keyword filtering heuristic and to label which dialect each document is written in, we hire Arabic annotators on MTurk to perform classification similar to Zaidan and Callison-Burch (2011b). Annotators were asked to classify the filtered documents for being in MSA or in one of four regional dialects: *Egyptian, Levantine, Gulf/Iraqi* or *Maghrebi*, and were shown the map in Figure 1 to explain what regions each of the dialect labels corresponded to. We allowed an additional "General" dialect option for ambiguous documents. Unlike Zaidan and Callison-Burch, our classification was applied to whole documents (corresponding to a user online posting) instead of individual sentences. To perform quality control, we used a set of documents for which correct labels were known. We presented these 20% of the time, and eliminated workers who did not correctly classify them (2% of labels).

Dialect Classification HIT	$10,064
Sentence Segmentation HIT	$1,940
Translation HIT	$32,061
Total cost	$44,065
Num words translated	1,516,856
Cost per word	2.9 cents/word

Table 1: The total costs for the three MTurk subtasks involved with the creation of our Dialectal Arabic-English parallel corpus.

Identifying the dialect of a text snippet can be challenging in the absence of phonetic cues. We therefore required 3 classifications from different workers for every document, and accepted a dialect label if at least two of them agreed. The dialect distribution of the final output was: 43% Gulf/Iraqi, 28% Levantine, 11% Egyptian, and 16% could not be classified. MSA and the other labels accounted for 2%. We decided to translate only the Levantine and Egyptian documents, since the pool of MTurk workers contained virtually no workers from Iraq or the Gulf region.

3.2 Sentence Segmentation

Since the data we annotated was mostly user-generated informal web content, the existing punctuation was often insufficient to determine sentence boundaries. Since sentence boundaries are important for correct translation, we segmented passages into individual sentences using MTurk. We only required sentences longer than 15 words to be segmented, and allowed Turkers to split and rejoin at any point between the tokens. The instructions were simply to "divide the Arabic text into individual sentences, where you believe it would be appropriate to insert a period." We also used a set of correctly segmented passages for quality control, and scored Turkers using a metric based on the precision and recall of correct segmentation points. The rejection rate was 1.2%.

3.3 Translation to English

Following Zaidan and Callison-Burch (2011a), we hired non-professional translators on MTurk to translate the Levantine and Egyptian sentences into

	Sentence	**Arabic**	**English**
Data Set	**Pairs**	**Tokens**	**Tokens**
MSA-150MW	8.0M	151.4M	204.4M
Dialect-1500KW	180k	1,545,053	2,257,041
MSA-1300KW	71k	1,292,384	1,752,724
MSA-Web-Tune	6,163	145,260	184,185
MSA-Web-Test	5,454	136,396	172,357
Lev-Web-Tune	2,600	20,940	27,399
Lev-Web-Test	2,600	21,092	27,793
Egy-Web-Test	2,600	23,671	33,565
E-Facebook-Tune	3,351	25,130	34,753
E-Facebook-Test	3,188	25,011	34,244

Table 2: Statistics about the training/tuning/test datasets used in our experiments. The token counts are calculated before MADA segmentation.

English. Among several quality control measures, we rendered the Arabic sentences as images to prevent Turkers from simply copying the Arabic text into translation software. We still spot checked the translations against the output of Google Translate and Bing Translator. We also rejected gobbledygook garbage translations that have a high percentage of words not found in an English lexicon.

We quantified the quality of an individual Turker's translations in two ways: first by asking native Arabic speaker judges to score a sample of the Turker's translations, and second by inserting control sentences for which we have good reference translations and measuring the Turker's METEOR (Banerjee and Lavie, 2005) and BLEU-1 scores (Papineni et al., 2002).[2] The rejection rate of translation assignments was 5%. We promoted good translators to a restricted access "preferred worker queue". They were paid at a higher rate, and were required to translate control passages only 10% of the time as opposed to 20% for general Turkers, thus providing us with a higher translation yield for unseen data.

Worker turnout was initially slow, but increased quickly as our reputation for being reliable payers was established; workers started translating larger volumes and referring their acquaintances. We had 121 workers who each completed 20 or more translation assignments. We eventually reached and sustained a rate of 200k words of acceptable quality translated per week. Unlike Zaidan and Callison-Burch (2011a), who only translated 2,000 Urdu sentences, we translated sufficient volumes of Dialectal Arabic to train machine translation systems. In total, we had 1.1M words of Levantine and 380k words of Egyptian translated into English, corresponding to about 2.3M words on the English side.

Table 1 outlines the costs involved with creating our parallel corpus. The total cost was $44k, or $0.03/word – an order of magnitude cheaper than professional translation.

4 Experiments in Dialectal Arabic-English Machine Translation

We performed a set of experiments to contrast systems trained using our dialectal parallel corpus with systems trained on a (much larger) MSA-English parallel corpus. All experiments use the same methods for training, decoding and parameter tuning, and we only varied the corpora used for training, tuning and testing. The MT system we used is based on a phrase-based hierarchical model similar to that of Shen et al. (2008). We used GIZA++ (Och and Ney, 2003) to align sentences and extract hierarchical rules. The decoder used a log-linear model that combines the scores of multiple feature scores, including translation probabilities, smoothed lexical probabilities, a dependency tree language model, in addition to a trigram English language model. Additionally, we used 50,000 sparse, binary-valued source and target features based on Chiang et al. (2009). The English language model was trained on 7 billion words from the Gigaword and from a web crawl. The feature weights were tuned to maximize the BLEU score on a tuning set using the Expected-BLEU optimization procedure (Devlin, 2009).

The Dialectal Arabic side of our corpus consisted of 1.5M words (1.1M Levantine and 380k Egyptian). Table 2 gives statistics about the various train/tune/test splits we used in our experiments. Since the Egyptian set was so small, we split it only to training/test sets, opting not to have a tuning set. The MSA training data we used consisted of Arabic-English corpora totaling 150M tokens (Arabic side). The MSA train/tune/test sets were constructed for the DARPA GALE program.

We report translation quality in terms of BLEU

[2] BLEU-1 provided a more reliable correlation with human judgment in this case that the regular BLEU score (which uses n-gram orders $1, \ldots, 4$), given the limited size of the sample measured.

		Simple Segment		MADA Segment			
Training	**Tuning**	**BLEU**	**OOV**	**BLEU**	**OOV**	**ΔBLEU**	**ΔOOV**
		MSA-Web-Test					
MSA-150MW	**MSA-Web**	26.21	1.69%	27.85	0.48%	+1.64	-1.21%
MSA-1300KW		21.24	7.20%	25.23	1.95%	+3.99	-5.25%
		Egyptian-Web-Test					
Dialect-1500KW	**Levantine-Web**	18.55	6.31%	20.66	2.85%	+2.11	-3.46%
		Levantine-Web-Test					
Dialect-1500KW	**Levantine-Web**	17.00	6.22%	19.29	2.96%	+2.29	-3.26%

Table 3: Comparison of the effect of morphological segmentation when translating MSA web text and Dialectal Arabic web text. The morphological segmentation uniformly improves translation quality, but the improvements are more dramatic for MSA than for Dialectal Arabic when comparing similarly-sized training corpora.

Training	**Tuning**	**BLEU**	**OOV**	**BLEU**	**OOV**	**BLEU**	**OOV**
		Egyptian-Web-Test		**Levantine-Web-Test**		**MSA-Web-Test**	
MSA-150MW	**MSA-Web**	14.76	4.42%	11.83	5.53%	**27.85**	0.48%
MSA-150MW	**Lev-Web**	14.34	4.42%	12.29	5.53%	24.63	0.48%
MSA-150MW+Dial-1500KW		20.09	**2.04%**	19.11	**2.27%**	24.30	**0.45%**
Dialect-1500KW		**20.66**	2.85%	**19.29**	2.96%	15.53	3.70%
Egyptian-360KW		19.04	4.62%	11.21	9.00%	-	-
Levantine-360KW		14.05	7.11%	16.36	5.24%	-	-
Levantine-1100KW		17.79	4.83%	**19.29**	3.31%	-	-

Table 4: A comparison of translation quality of Egyptian, Levantine, and MSA web text, using various training corpora. The highest BLEU scores are achieved using the full set of dialectal data (which combines Levantine and Egyptian), since the Egyptian alone is sparse. For Levantine, adding Egyptian has no effect. In both cases, adding MSA to the dialectal data results in marginally worse translations.

score.[3] In addition, we also report the OOV rate of the test set relative to the training corpus in each experimental setups.

4.1 Morphological Decomposition

Arabic has a complex morphology compared to English. Preprocessing the Arabic source by morphological segmentation has been shown to improve the performance of Arabic MT (Lee, 2004; Habash and Sadat, 2006) by decreasing the size of the source vocabulary, and improving the quality of word alignments. The morphological analyzers that underlie most segmenters were developed for MSA, but the different dialects of Arabic share many of the morphological affixes of MSA, and it is therefore not unreasonable to expect MSA segmentation to also improve Dialect Arabic to English MT. To test this, we ran experiments using the MADA morphological analyzer (Habash and Rambow, 2005). Table 3 shows the effect of applying segmentation to the text, for both MSA and Dialectal Arabic. The BLEU score improves uniformly, although the improvements are most dramatic for smaller datasets, which is consistent with previous work (Habash and Sadat, 2006). Morphological segmentation gives a smaller gain on dialectal input, which could be due to two factors: the segmentation accuracy likely decreases since we are using an unmodified MSA segmenter, and there is higher variability in the written form of dialect compared to MSA. Given the significant, albeit smaller gain on dialectal input, we use MADA segmentation in all our experiments.

4.2 Effect of Dialectal Training Data Size

We next examine how the size of the dialectal training data affects MT performance, and whether it is useful to combine it with MSA training data. We

[3] We also computed TER (Snover et al., 2006) and METEOR scores, but omit them because they demonstrated similar trends.

Arabic	TL	Count	English Equivalent
عنجد	*Enjd*	31	really/for real – Levantine.
كتيير	*ktyyyr*	17	a looot (corruption of MSA *kvyrA*).
النعوم	*AlnEwm*	16	The last name (Al-Na'oom) of a forum admin.
وحشتينى	*wH$tyny*	14	I miss you (spoken to a female) – Egyptian.
يازمن	*yAzmn*	11	oh time (space omitted). Appeared within a poem.
بكتير	*bktyr*	11	by much (corruption of MSA *bkvyr*).
متلك	*mtlk*	10	like you (corruption of MSA *mvlk*).

Table 5: The most frequent OOV's (with counts ≥ 10) of our test sets against the MSA training data.

Source (*EGY*):	انت بتعمل له اعلان ولا ايه؟!!
Transliteration:	*Ant btEml lh AElAn wlA Ayh?!!*
MSA-Sys. Output:	You are working for a declaration and not?
Dial-Sys. Output:	You are making the advertisement for him or what?
Reference:	Are you promoting it or what?!!
Source (*EGY*):	نفسي اطمئن عليه بعد ما شاف الصوره دي
Transliteration:	*nfsY Atm}n Elyh bEd mA $Af AlSwrh dy*
MSA-Sys. Output:	Myself feel to see this image.
Dial-Sys. Output:	I wish to check on him after he saw this picture.
Reference:	I wish to be sure that he is fine after he saw this images
Source (*LEV*):	لهيك الجو كتيير كووول
Transliteration:	*lhyk Aljw ktyyyr kwwwl*
MSA-Sys. Output:	God you the atmosphere.
Dial-Sys. Output:	this is why the weather is so cool
Reference:	This is why the weather is so cool
Source (*LEV*):	طول بالك عم نمزح
Transliteration:	*Twl bAlk Em nmzH*
MSA-Sys. Output:	Do you think about a joke long.
Dial-Sys. Output:	Calm down we are kidding
Reference:	calm down, we are kidding

Figure 2: Examples of improvement in MT output when training on our Dialectal Arabic-English parallel corpus instead of an MSA-English parallel corpus.

Source (*EGY*):	قالتلة **طب** تعالى نعد ،
Transliteration:	*qAltlp **Tb** tEAlY nEd ,*
MSA-Sys. Output:	**Medicine** almighty promise.
Dial-Sys. Output:	She said, **OK**, come and then
Reference:	She told him, **OK**, lets count them ,
Source (*LEV*):	فبقرا وأحيانا بقضيها **عم** أتسلى مع رفقاتي
Transliteration:	*fbqrA w>HyAnA bqDyhA **Em** >tslY mE rfqAty*
MSA-Sys. Output:	I read and sometimes with go with my **uncle**.
Dial-Sys. Output:	So I read, and sometimes I spend try**ing** to make my self comfort with my friends
Reference:	So i study and sometimes I spend the time hav**ing** fun with my friends
Source (*LEV*):	الله يسامحكن هلق كل واحد طالب قرب بيكون **بدو** عروس
Transliteration:	*Allh ysAmHkn hlq kl wAHd TAlb qrb bykwn **bdw** Erws*
MSA-Sys. Output:	God now each student near the **Bedouin** bride.
Dial-Sys. Output:	God forgive you, each one is a close student would **want** the bride
Reference:	God forgive you. Is every one asking to be close, **want** a bride!

Figure 3: Examples of ambiguous words that are translated incorrectly by the MSA-English system, but correctly by the Dialectal Arabic-English system.

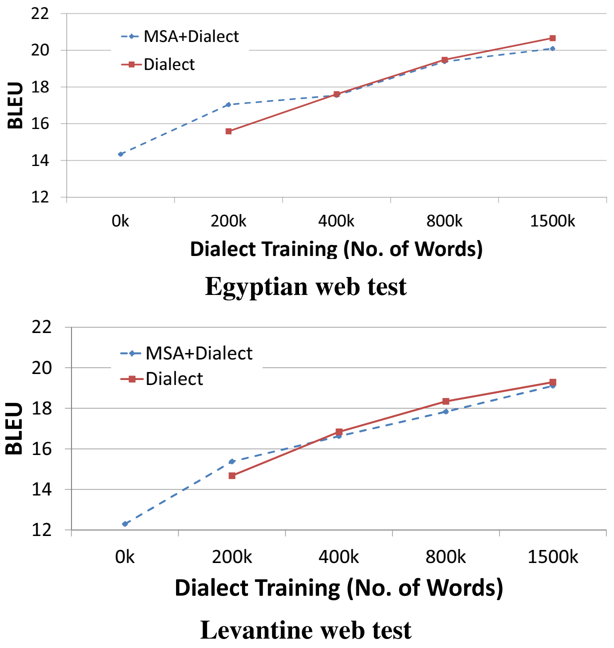

Figure 4: Learning curves showing the effects of increasing the size of dialectal training data, when combined with the 150M-word MSA parallel corpus, and when used alone. Adding the MSA training data is only useful when the dialectal data is scarce (200k words).

started with a baseline system trained on the 150M-word MSA parallel corpus, and added various sized portions of the dialect parallel corpus to it. Figure 4 shows the resulting learning curve, and compares it to the learning curve for a system trained solely on the dialectal parallel corpus. When only 200k words of dialectal data are available, combining it with the 150M-word MSA corpus results in improved BLEU scores, adding 0.8–1.5 BLEU points. When 400k words or more of dialectal data are available, the MSA training data ceases to provide any gain, and in fact starts to hurt the performance.

The performance of a system trained on the 1.5M-word dialectal data is dramatically superior to a system that uses only the 150M-word MSA data: +6.32 BLEU points on the Egyptian test set, or 44% relative gain, and +7.00 BLEU points on the Levantine test set, or 57% relative gain (fourth line vs. second line of Table 4). In Section 4.4, we show that those gains are not an artifact of the similarity between test and training datasets, or of using the same translator pool to translate both sets.

Inspecting the difference in the outputs of the Dialectal vs. MSA systems, we see that the improvement in score is a reflection of a significant improvement in the quality of translations. Figure 2 shows a few examples of sentences whose translations improve significantly using the Dialectal system. Figure 3 shows a particularly interesting category of examples. Many words are homographs, with different meanings (and usually different pronunciations) in MSA vs. one or more dialects. The bolded tokens in the sentences in Figure 3 are examples of such words. They are translated incorrectly by the MSA system, while the dialect system translates them correctly.[4] If we examine the most frequent OOV words against the MSA training data (Table 5), we find a number of corrupted MSA words and names, but that a majority of OOVs are dialect words.

4.3 Cross-Dialect Training

Since MSA training data appeared to have little effect when translating dialectal input, we next investigated the effect of training data from one dialect on translating the input of another dialect. We trained a system with the 360k-word Egyptian training subset of our dialectal parallel corpus, and another system with a similar amount of Levantine training data. We used each system to translate the test set of the other dialect. As expected, a system performs better when it translates a test set in the same dialect that it was trained on (Table 4).

That said, since the Egyptian training set is so small, *adding* the (full) Levantine training data improves performance (on the Egyptian test set) by 1.62 BLEU points, compared to using only Egyptian training data. In fact, using the Levantine training data by itself outperforms the MSA-trained system on the Egyptian test set by more than 3 BLEU points. (For the Levantine test set, adding the Egyptian training data has no affect, possibly due to the small amount of Egyptian data.) This may suggest that the mismatch between dialects is less severe than the mismatch between MSA and dialects. Alternatively, the differences may be due to the changes in genre from the MSA parallel corpus (which is mainly formal newswire) to the newsgroups and weblogs that mainly comprise the dialectal corpus.

[4] The word *nfsY* of Figure 2 (first word of second example) is also a homograph, as it means *myself* in MSA and *I wish* in Dialectal Arabic.

Training	**Tuning**	**BLEU**	**OOV**
MSA-150MW	**Levantine-Web**	13.80	4.16%
MSA-150MW+Dialect-1500KW		**16.71**	**2.43**%
Dialect-1500KW		15.75	3.79%
MSA-150MW	**Egyptian-Facebook**	15.80	4.16%
MSA-150MW+Dialect-1500KW		**18.50**	**2.43**%
Dialect-1500KW		17.90	3.79%
Dialect-1000KW (random selection)	**Egyptian-Facebook**	17.09	4.64%
Dialect-1000KW (no Turker overlap)		17.10	4.60%

Table 6: Results on a truly independent test set, consisting of data harvested from Egyptian Facebook pages that are entirely distinct from the our dialectal training set. The improvements over the MSA baseline are still considerable: +2.9 BLEU points when no Facebook data is available for tuning and +2.7 with a Facebook tuning set.

4.4 Validation on Independent Test Data

To eliminate the possibility that the gains are solely due to similarity between the test/training sets in the dialectal data, we ran experiments using the same dialectal training data, but using truly independent test/tuning data sets selected at random from a larger set of monolingual data that we collected from public Egyptian Facebook pages. This data consists of a set of original user postings and the subsequent comments on each, giving the data a more conversational style than our other test sets. The postings deal with current Egyptian political affairs, sports and other topics. The test set we selected consisted of 25,011 words (3,188 comments and 427 postings from 86 pages), and the tuning set contained 25,130 words (3,351 comments and 415 conversations from 58 pages). We obtained reference translations for those using MTurk as well.

Table 6 shows that using the 1.5M-word dialect parallel corpus for training yields a 2 point BLEU improvement over using the 150M-word MSA corpus. Adding the MSA training data does yield an improvement, though of less than a single BLEU point. It remains true that training on 1.5M words of dialectal data is better than training on 100 times more MSA parallel data. The system performance is sensitive to the tuning set choice, and improves when it matches the test set in genre and origin.

To eliminate another potential source of artificial bias, we also performed an experiment where we removed any training translation contributed by a Turker who translated any sentence in the Egyptian Facebook set, to eliminate translator bias. For this, we were left with 1M words of dialect training data. This gave the same BLEU score as when training with a randomly selected subset of the same size (bottom part of Table 6).

4.5 Mapping from Dialectal Arabic to MSA Before Translating to English

Given the large amount of linguistic resources that have been developed for MSA over the past years, and the extensive research that was conducted on machine translation from MSA to English and other languages, an obvious research question is whether Dialectal Arabic is best translated to English by first pivoting through MSA, rather than directly. The proximity of Dialectal Arabic to MSA makes the mapping in principle easier than general machine translation, and a number of researchers have explored this direction (Salloum and Habash, 2011). In this scenario, the dialectal source would first be automatically transformed to MSA, using either a rule-based or statistical mapping module.

The Dialectal Arabic-English parallel corpus we created presents a unique opportunity to compare the MSA-pivoting approach against direct translation. First, we collected equivalent MSA data for the Levantine Web test and tuning sets, by asking Turkers to transform dialectal passages to valid and fluent MSA. Turkers were shown example transformations, and we encouraged fewer changes where applicable (e.g. morphological rather than lexical mapping), but allowed any editing operation in general (deletion, substitution, reordering). Sample submissions were independently shown to native Arabic speaking judges, who confirmed they were valid MSA. A low OOV rate also indicated the correctness of the mappings. By manually transforming the test

Training	**BLEU**	**OOV**	**BLEU**	**OOV**	**∆BLEU**	**∆OOV**
	Direct dialect trans		**Map to MSA then trans**			
MSA-150MW	12.29	5.53%	14.59	1.53%	+2.30	-4.00%
MSA-150MW+Dialect-200KW	15.37	3.59%	15.53	1.22%	+0.16	-2.37%
MSA-150MW+Dialect-400KW	16.62	3.06%	16.25	1.13%	-0.37	-1.93%
MSA-150MW+Dialect-800KW	17.83	2.63%	16.69	1.04%	-1.14	-1.59%
MSA-150MW+Dialect-1500KW	19.11	2.27%	17.20	0.98%	-1.91	-1.29%

Table 7: A comparison of the effectiveness of performing Levantine-to-MSA mapping before translating into English, versus translating directly from Levantine into English. The mapping from Levantine to MSA was done manually, so it is an optimistic estimate of what might be done automatically. Although initially helpful to the MSA baseline system, the usefulness of pivoting through MSA drops as more dialectal data is added, eventually hurting performance.

dialectal sentence into MSA, we establish an optimistic estimate of what could be done automatically.

Table 7 compares direct translation versus pivoting to MSA before translating, using the baseline MSA-English MT system.[5] The performance of the system improves by 2.3 BLEU points with dialect-to-MSA pivoting, compared to attempting to translate the untransformed dialectal input directly. As we add more dialectal training data, the BLEU score when translating the untransformed dialect test set improves rapidly (as seen previously in the MSA+Dialect learning curve in Figure 4), while the improvement is less rapid when the text is first transformed to MSA. Direct translation becomes a better option than mapping to MSA once 400k words of dialectal data are added, despite the significantly lower OOV rate with MSA-mapping. This indicates that simple vocabulary coverage is not sufficient, and data domain mismatch, quantified by more complex matching patterns, is more important.

5 Conclusion

We have described a process for building a Dialectal Arabic-English parallel corpus, by selecting passages with a relatively high percentage of non-MSA words from a monolingual Arabic web text corpus, then using crowdsourcing to classify them by dialect, segment them into individual sentences and translate them to English. The process was successfully scaled to the point of reaching and sustaining a rate of 200k translated words per week, at $1/10$ the cost of professional translation. Our parallel corpus, consisting of 1.5M words, was produced at a total cost of \$40k, or roughly \$0.03/word.

We used the parallel corpus we constructed to analyze the behavior of a Dialectal Arabic-English MT system as a function of the size of the dialectal training corpus. We showed that relatively small amounts of training data render larger MSA corpora from different data genres largely ineffective for this test data. In practice, a system trained on the combined Dialectal-MSA data is likely to give the best performance, since informal Arabic data is usually a mixture of Dialectal Arabic and MSA. An area of future research is using the output of a dialect classifier, or other features to bias the translation model towards the Dialectal or the MSA parts of the data.

We also validated the models built from the dialectal corpus by using them to translate an independent data set collected from Egyptian Facebook public pages. We finally investigated using MSA as a "pivot language" for Dialectal Arabic-English translation, by simulating automatic dialect-to-MSA mapping using MTurk. We obtained limited gains from mapping the input to MSA, even when the mapping is of good quality, and only at lower training set sizes. This suggests that the mismatch between training and test data is an important aspect of the problem, beyond simple vocabulary coverage.

The aim of this paper is to contribute to setting the direction of future research on Dialectal Arabic MT. The gains we observed from using MSA morphological segmentation can be further increased with dialect-specific segmenters. Input preprocessing can also be used to decrease the noise of the user-generated data. Topic adaptation is another important problem to tackle if the large MSA linguistic resources already developed are to be leveraged for Dialectal Arabic-English MT.

[5] The systems in each column of the table are tuned consistently, using their corresponding tuning sets.

Acknowledgments

This work was supported in part by DARPA/IPTO Contract No. HR0011-12-C-0014 under the BOLT Program, and in part by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Distribution Statement A (Approved for Public Release, Distribution Unlimited).

References

Hitham M. 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, Egypt.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *In Proc. of ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, June.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *NAACL '09: Proceedings of the 2009 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado.

Jacob Devlin. 2009. Lexical features for statistical machine translation. Master's thesis, University of Maryland, December.

Mohamed Embarki and Moha Ennaji, editors. 2011. *Modern Trends in Arabic Dialectology*. The Red Sea Press.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15:325–340.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sydney, Australia.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, New York.

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

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL '04: Proceedings of HLT-NAACL 2004*, Boston, Massachusetts.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):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 the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA.

Jason Riesa and David Yarowsky. 2006. Minimally supervised morphological segmentation with applications to machine translation. In *Proceedings of the 7th Conf. of the Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, MA.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the 2011 Conference of Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conf. of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 577–585, Columbus, Ohio.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the 7th Conf. of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, MA.

Omar F. Zaidan and Chris Callison-Burch. 2011a. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, June.

Omar F. Zaidan and Chris Callison-Burch. 2011b. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, June.