Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation

Kamla Al-Mannai[1], **Hassan Sajjad**[1], **Alaa Khader**[2], **Fahad Al Obaidli**[1], **Preslav Nakov**[1], **Stephan Vogel**[1]
Qatar Computing Research Institute[1], Carnegie Mellon University in Qatar[2]
{kamlmannai,hsajjad,faalobaidli,pnakov,svogel}@qf.org.qa[1], akhader@cmu.edu[2]

Abstract

We demonstrate the feasibility of using unsupervised morphological segmentation for dialects of Arabic, which are poor in linguistics resources. Our experiments using a Qatari Arabic to English machine translation system show that unsupervised segmentation helps to improve the translation quality as compared to using no segmentation or to using ATB segmentation, which was especially designed for Modern Standard Arabic (MSA). We use MSA and other dialects to improve Qatari Arabic to English machine translation, and we show that a uniform segmentation scheme across them yields an improvement of 1.5 BLEU points over using no segmentation.

1 Introduction

The Arabic language has many varieties, where the Modern Standard Arabic (MSA) coexists with various dialects. Dialects differ from MSA and from each other lexically, phonologically, morphologically and syntactically. MSA has standard orthography and is used in formal contexts (e.g., publications, newspaper articles, etc.), while the dialects are usually limited to daily verbal interactions. However, with the recent rise of social media, it has become increasingly common to use dialects in written communication as well, which has constituted the research in dialectal Arabic (DA) as a separate field within the broader field of natural language processing (NLP).

As DA NLP is still in its infancy, there is lack of basic computational resources and tools, which are needed in order to apply standard NLP approaches to the dialects of Arabic. For instance, statistical approaches need a lot of training data, which makes it very hard, if not impossible, to apply them to resource-poor languages; this is especially true for statistical machine translation (SMT) of Arabic dialects.

The Arabic language and its dialects are highly inflectional, and a word can appear in many more inflected forms compared to English. Consider the Arabic words لعبت ,يلعب ,تلعب, and يلعبون: they all belong to one root word لعب ‘playing’ /lEb/. Each morphological variation is derived from a root word with different affixes addressing different functions. This causes data sparseness, and covering all possible word forms of a root word may not be always possible. Considering the different variants of Arabic, the problem is exacerabated as dialects could use different choices of affixes for the same function. For example, the MSA word يلعبون /yalEabuwn/, meaning ‘they are playing’, could be found as يلعبون /ylEbuwn/ in Gulf, as عم يلعبوا /Eam yilEabuA/ in Levantine, and as بيلعبوا /biylEabwA/ in Egyptian Arabic.

One possible solution is to use a morphological segmenter that segments words into simpler units such as stems and affixes, which might be covered in the training set (Zollmann et al., 2006; Tsai et al., 2010). When applied to dialects, this may reduce the lexical gap between dialects and MSA by matching the common stems. Unfortunately, there are no standard morphological segmentation tools for dialects. Due to the difference in morphology, tools designed for MSA do not work well for dialects. Developing rule-based segmenters for each dialect might appear to be the ideal solution, but, as the orthography of dialects is not standardized, crafting linguistic rules for them is very hard.

In this paper, we focus on training an unsupervised model for word segmentation, which we apply to SMT for a given Arabic dialect. We train a pre-existing unsupervised segmentation model on the Arabic side of the training bi-text (and on some other monolingual data), and then we optimize its parameters based on the resulting SMT quality. Similarly, a *multi-dialectal* word segmenter could be developed by training on multi-dialectal data.

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 207–216,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

In particular, we develop a Qatari Arabic to English (QA-EN) SMT system, which we train on a small pre-existing bi-text. As part of the development of the unsupervised segmentation model, we also collected some additional monolingual data for Qatari Arabic. Qatari Arabic is a subdialect of the more general Gulf dialect, among with Saudi, Kuwaiti, Emirati, Bahraini, and Omani; we collected additional monologual data for each of these subdialects, and we release this data to the research community.

We train an unsupervised segmentation tool, Morphessor, and its MAP model (Creutz and Lagus, 2007), using different variations of the collected Qatari data. We optimize the single hyperparameter of the MAP model by maximizing the translation quality of the QA-EN SMT system in terms of BLEU. Our experimental results demonstrate that the resulting unsupervised segmenter yields improvements in translation quality when compared to (i) using no segmentation and (ii) using an MSA-based ATB segmenter.

We further develop a multi-dialectal word segmentation model, which we train on the Arabic side of the multi-dialectal training data, which consists of Qatari Arabic, Egyptian Arabic (EGY), Levantine Arabic (LEV) and MSA to English, i.e., a scaled combination of all the available parallel data. We train a QA-EN SMT system using the segmented multi-dialectal data, and we show an absolute gain of 1.5 BLEU points compared to a baseline that uses no segmentation.

The rest of the paper is organized as follows: First, we provide an overview of related work on Dialectal Arabic NLP (Section 2). Next, we discuss and we illustrate the linguistic differences between different Arabic dialects in comparison with and with a focus on Qatari Arabic (Section 3). Then, we provide statistics about the corpora we collected and used in our experiments, followed by an illustration of the orthographic normalization schemes we applied (Section 4). We next provide a high-level description of our approach, which uses morphological segmentation to combine resources for other Arabic dialects in a QA-EN SMT system effectively (Section 4.3). We also explain our experimental setup and we present the results (Section 5). We then discuss translating in the reverse direction, i.e., into Qatari Arabic (Section 6). Finally, we point to possible directions for future work and we conclude the paper (Section 7).

2 Related Work

NLP for DA is still in its early stages of development and many challenges need to be overcomed such as the lack of suitable tools and resources.

Collecting resources for dialectal Arabic: Several researchers have directed efforts to develop DA computational resources (Maamouri et al., 2006; Al-Sabbagh and Girju, 2010; Zaidan and Callison-Burch, 2011; Salama et al., 2014). Zbib et al. (2012) built two dialectal Arabic-English parallel corpora for Egyptian and Levantine Arabic using crowdsourcing. Bouamor et al. (2014) presented a multi-dialectal Arabic parallel corpus, which covers five Arabic dialects besides MSA and English. Mubarak and Darwish (2014) collected a multi-dialectal corpus using Twitter. Unlike previous work, we focus on Gulf subdialects, particularly Qatari Arabic. The monolingual data that we collected is a high-quality dialectal resource and originates from dialect-specific sources such as novels and forums.

Adapting SMT resources for other Arabic dialects: Many researchers have explored the potential of using MSA as a pivot language for improving SMT of Arabic dialects (Bakr et al., 2008; Sawaf, 2010; Salloum and Habash, 2011; Sajjad et al., 2013a; Jeblee et al., 2014). This often involves DA-MSA conversion schemes as an alternative in the absence of DA-MSA parallel resources. In contrast, limited work has been done on leveraging available resources for other dialects. Recently, Zbib et al. (2012) have shown that using a small amount of dialectal data could yield great improvements for SMT. Here, we investigate the potential of improving the resource adaptability of Arabic dialects. Our work is different as we use an unsupervised segmenter that helps in improving the lexical overlap between dialects and MSA.

Building morphological segmenters for the Arabic dialects: Researchers have already focused efforts on crafting and extending existing MSA tools to DA by mainly using a set of rules (Habash et al., 2012). Habash and Rambow (2006) presented MAGEAD, a knowledge-based morphological analyzer and generator for Egyptian and Levantine Arabic. Chiang et al. (2006) developed a Levantine morphological analyzer on top of an existing MSA analyzer using an explicit knowledge base.

Riesa and Yarowsky (2006) trained a supervised trie-based model using a small lexicon of dialectal affixes. In our work, we eliminate the need for linguistic knowledge by training an unsupervised model using available resources. The unsupervised mode of learning allowed us to develop a multi-dialectal morphological segmenter.

3 Arabic Dialects

In this section, we highlight some of the linguistic differences between Arabic dialects and MSA, with a focus on the Qatari dialect.

3.1 Phonological Variations

The Gulf dialect often preserves the phonological representation of MSA, which is not the case with many other Arabic dialects. For example, in Egyptian (EGY) and in some Levantine (LEV) dialects, the MSA consonants ث /v/, ق /q/, and ذ /*/ are realized as ت /t/, glottal stop /'/, and ظ /Z/, respectively. While, their MSA pronunciations are preserved in Gulf Arabic.

In Gulf Arabic, there are some phonological differences between countries such as Kuwait (KW), Saudi Arabia (SA), Bahrain (BH), Qatar (QA), United Arab Emirates (AE), and Oman (OM). Here, we focus our discussion on Qatari Arabic, and we compare it to MSA and other dialects.

The QA dialect borrows two Persian characters namely چ /J/ and ڤ /V/. For instance, the MSA letter ج /j/ is converted to /J/ in QA, e.g., إجتماع 'meeting' is pronounced as /<jtimAE/ in MSA and /<JtimAE/ in QA. The Persian character چ /J/ is also used in place of ك /k/ in some MSA words when they are used in QA. For example, سمك 'fish' /samak/ is pronounced سمچ /smaJ/ in QA, while the EGY and the LEV dialects maintain the MSA pronunciation. The Persian ڤ /V/ is used to map the sound of the English letter 'v' in borrowed foreign words, e.g., فيديو 'video' is pronounced as ڤيديو /Viydyw/ as opposed to /fiydywu/; the form in which it is written in MSA.

The MSA consonant ض /D/ is not used in the QA dialect. It is substituted by ظ /Z/ in Qatari. For example, the MSA pronunciation /HaD/ of حض 'to encourage' is transformed to حظ /HaZ/ in QA, but it is maintained in EGY.

Meanwhile, the MSA consonant ظ /Z/ is realized as /D/ in EGY. For example, the MSA pronunciation /HaZ/ of حظ 'luck' is maintained in QA and transformed to /HaD/ in EGY. This change is consistent in all words within each dialect. However, such phonological variations between dialects have the potential to add ambiguity to dialectal Arabic.

The MSA consonant ج /j/ can be used to distinguish between different dialects, particularly Gulf subdialects. ج /j/ is pronounced as ي /y/ in KW, BH, QA, AE, ق /q/ in OM, much like in EGY, and ج /j/ in SA, much like in LEV. For example, the MSA word مسجد 'mosque' /masjid/ is pronounced as /masjid/ in MSA, SA, LEV, مسقد /masqid/ in OM, EGY, مسيد /masyid/ in KW, BH, QA, AE, while the MSA pronunciation is preserved in SA. This change does not apply to names. However, we should note that it is not consistent in QA, e.g., the MSA pronunciation of ج /j/ in جبل 'mountain' /jabal/ and برج 'tower' /burj/ is preserved in QA.

3.2 Morphological Variations

In Arabic, a root can produce surface wordforms by means of inflectional and derivational morphological processes (Habash, 2010).

An inflectional word form is a variant of a root word with the same meaning but expressing a different function, e.g., gender, number, case. It is usually formed by adding a prefix, a suffix, or a circumfix to a stem word. Note that Arabic dialects can make different lexical choices for affixations compared to MSA. For example, the MSA future prefix س /s/ is replaced by ب /b/ in QA and by هـ /h/ in EGY and LEV. Thus, the MSA word سيأكل 'he will eat' /say>kul/ becomes بياكل /biyAkil/ in QA and هياكل /hayAkul/ in EGY and LEV.

A derivational word form is formed by applying a pattern to a root word, e.g., 'player' is derived from 'play' using the pattern `noun + 'er'`. An example of an Arabic derivational form is تفعل 'do' /tafaEãl/. The root is فعل /faEal/ and it uses the imperative pattern فعل+ت. In EGY, ا /A/ is added as a prefix; so, it becomes اتفعل /AitfaEĩl/.

Meanwhile, the original form is preserved in QA.

Changing the structure of a pattern in a dialect will result in producing a new dialect-specific orthography for every word that is represented by the structure. For example, the MSA word تعلم 'learn' /taEalãm/ becomes اتعلم /AitEalim/ in EGY, while the MSA form is preserved in QA.

3.3 Lexical Variations

Lexical variations are among the most obvious differences between Arabic dialects. For example, the MSA word ماذا 'what' /mA*A/ would be found as شو /$uw/ in LEV, إيه /<yh/ in EGY, and شنو /$nuw/ in GLF. We can find lexical variations in subdialects as well. For example, the MSA negation word لن /lan/, 'not', is expressed as مب /mab/ in QA, as مو /muw/ in KW, and as مهب /ma-hab/ in SA.

3.4 Orthographic Variations

Due to the lack of orthographic standardization of dialectal Arabic, some MSA words can be found in dialectal text with both MSA and phonological spellings. For example, the MSA word جمعة 'gathering' /jamEap/ can be also spelled as يمعه /yamEah/, which is a phonetic variation in QA. Some dialectal words also vary in spelling due to variation in their pronunciation, e.g., أشوف /A$uwf/, a QA word meaning 'I see', can be also spelled as اجوف /Ajuwf/.

In dialectal Arabic, different orthographic forms are also possible for entire phrases. For instance, words followed or preceded by pronouns are commonly reduced to a single word, e.g., قلت لها /glt lahA/ 'I told her' is written as قلتلها. Also, commonly used religious phrases can be found written as a single unit, e.g., ما شاء الله /mA $A' Aĩah/ 'God has willed it' as مشالله.

4 Methodology

In the section, we present some statistics about the Arabic dialectal data that we have collected. We processed it to remove orthographic inconsistencies. Then, we used a pre-existing unsupervised morphological segmenter, Morfessor, in order to segment the text.

Corpus	**QCA**	**AVIA$_{QA}$**	**AVIA$_{O}$**
Sents	14.7	0.9	2
Tokens	115	6.7	15

Table 1: Statistics about the collected parallel corpora (in thousands). AVIA$_O$ shows the statistics about the AVIA corpus excluding Qatari data.

4.1 Data Collection

We did an extensive search for available monolingual and bilingual resources for the Gulf dialect, with a focus on Qatari Arabic. Tables 1 and 2 present some statistics about the corpora we collected. More detailed description follows below.

Bilingual corpora:

– The **QCA speech corpus**, comprises 14.7k sentences that are phonetically transcribed from TV broadcasts in Qatari Arabic and translated to English; see (Elmahdy et al., 2014) for more detail. The corpus was designed for speech recognition and we faced several normalization-related issues that we had to resolve before it could be used for machine translation and language modeling. One example is the usage of five Persian characters to represent some sounds in Arabic words. Moreover, the English side had some grammatical and spelling errors. We normalized the Arabic side and corrected the English side of the corpus as described in Section 4.2. The corpus can be found at `http://sprosig.isle.illinois.edu/corpora/1`.

– The **AVIA corpus**[1] is designed as a reference source of dialectal Arabic. It consists of 3k sentences in four Gulf subdialects: Emirati (AE), Kuwaiti (KW), Qatari (QA), and Hejazi (SA).[2] The data consists of dialectal sentences that contain words commonly used in daily conversation.

Monolingual corpora: We further collected monolingual corpora consisting of a total of 2.7M tokens for various Gulf subdialects. The Qatari part of the data consists of 470K tokens. Most of the corpus is a collection of novels, belonging to the romance genre.[3] For the Qatari dialect, we also collected Qatari forum data.[4]

[1] `http://terpconnect.umd.edu/~nlynn/AVIA/Level3/`

[2] The website also contains small parallel corpora for MSA, EGY and LEV to English, but here we focus on Gulf subdialects only.

[3] `http://forum.te3p.com/264311-52.html`

[4] `www.qatarshares.com/vb/index.php`

Corpus	Novel						Forum
	AE	BH	KW	OM	QA	SA	QA
Tokens	573	244	178	372	412	614	69
Types	43	22	27	27	43	71	15

Table 2: Statistics about the collected monolingual corpora (in thousands of words).

To the best of our knowledge, this is the first collection of monolingual corpora for Gulf Arabic subdialects. It can be helpful for, e.g., language modeling when translating into Arabic, for learning the similarities and differences between Gulf subdialects, etc. Table 2 shows some statistics about the data after punctuation tokenization.

4.2 Orthographic Normalization

The inconsistency in the orthographic spelling of the same word can increase data sparseness. Thus, we normalize the Arabic text in the collected resources by applying the reduced orthographic normalization scheme, e.g., Tah Marbota is reduced to Hah. We also normalize extended lines between letters, e.g., ســـكـــر 'sugar' /sukar/ is changed to سكر, and we reduce character elongations to be just two characters long. In order to maintain consistency among different resources, we remove supplementary diacritics, e.g., عُقَدْ 'knots' /Euqad/ is normalized to عقد, and we map Persian letters to their phonological correspondences in Qatari Arabic[5], i.e., گ /G/ to ق /g/, ڤ /V/ to ف /f/, پ /P/ to ب /b/, and ژ and چ /J/ to ج /j/.

For the English texts, the orthographic variations were already normalized. However, the English side of the QCA corpus had some spelling and grammatical errors, which we corrected manually. On the grammatical side, we only corrected a subset of the data, which we used for tuning and testing our SMT system (see Section 5).

4.3 Morphological Decomposition

There is no general Arabic morphological segmenter that works for all variations of Arabic. The most commonly used segmenters for Arabic were designed for MSA (Habash et al., 2009; Green and DeNero, 2012). Due to the lexical and morphological differences between dialects and MSA, these MSA-based morphological tools do not work well for dialects.

In this work, we used an unsupervised morphological segmenter, Morfessor-categories MAP[6], an unsupervised model with a single hyperparameter (Creutz and Lagus, 2007). We chose Morfessor because of its superior performance on Arabic compared to other unsupervised models (Siivola et al., 2007; Poon et al., 2009).

The model has a single hyperparameter, the perplexity threshold parameter B, which controls the granularity of segmentation. The recommended value ranges from 1 to 400 where 1 means maximum fine-grained segmentation, and 400 restricts it to the least segmented output. We set the threshold empirically to 70, as shown in Section 5.1.

5 Experimental Setup

We performed an extrinsic evaluation of the variations in segmentation by building a Qatari Arabic to English machine translation system on each of them. We also tested Morfessor on other available dialects and on MSA, and we will show below how a uniform segmentation can help to better adapt resources for dialects and MSA for SMT. This section describes our experimental setup.

Datasets: We divided the **QCA corpus** into 1k sentences each for development and testing, and we used the remaining 12k for training.

We adapted parallel corpora for **Egyptian**, **Levantine** and **MSA** to English to be used for Qatari Arabic to English SMT. For MSA, we used parallel corpora of TED talks (Cettolo et al., 2012) and the AMARA corpus (Abdelali et al., 2014), which consists of educational videos. Since the QCA corpus is in the speech domain, we believe that an MSA corpus of spoken domain would be more helpful than a text domain such as News. For Egyptian and Levantine, we used the parallel corpus provided by Zbib et al. (2012). There is no Gulf–English parallel data available in the literature. The data that we found was a very small collection of subdialects of Gulf Arabic; we did not use it for MT experiments. However, we used the Qatari part of the AVIA corpus to train Morfessor.

Machine translation system settings: We used a phrase-based statistical machine translation model as implemented in the Moses toolkit (Koehn et al., 2007) for machine translation.

[5] This issue relates to the QCA corpus.

[6] This is an extension of the basic Morfessor method and is based on a Maximum a Posteriori model.

We built separate directed word alignments for source-to-target and target-to-source using IBM model 4 (Brown et al., 1993), and we symmetrized them using the grow-diag-final-and heuristics (Koehn et al., 2003). We then extracted phrase pairs with a maximum length of seven, and we scored them using maximum likelihood estimation with Kneser-Ney smoothing (Kneser and Ney, 1995). We also built a lexicalized reordering model, msd-bidirectional-fe. We built a 5-gram language model on the English side of QCA-train using KenLM (Heafield, 2011). Finally, we built a log-linear model using the above features.

We tuned the model weights by optimizing BLEU (Papineni et al., 2002) on the tuning set, using PRO (Hopkins and May, 2011) with sentence-level BLEU+1 optimization (Nakov et al., 2012). In testing, we used minimum Bayes risk decoding (Kumar and Byrne, 2004), cube pruning, and the operation sequence model (Durrani et al., 2011).

Baseline: Our baseline Qatari Arabic to English MT system is trained on the QCA bitext without any segmentation of Qatari Arabic. For the experiments described in this paper, we used the English side of the QCA corpus for language modeling.

5.1 Experimental Results

In this section, we first present our work on using Morfessor for segmenting Qatari Arabic. We tried different values of its parameter, and we trained it using corpora of different sizes to find balanced settings that improve SMT quality as compared with no segmentation and with segmentation using the Stanford ATB segmenter. We further applied our selected settings to segment MSA, EGY and LEV and used them for Qatari Arabic to English machine translation. Our results show that a uniform segmentation scheme across different dialects improves machine translation.

Morfessor training variations: We trained Morfessor using three corpora: (i) QCA, (ii) AVIA$_{QA}$ plus Qatari Novels, and (iii) a combination thereof. Table 3 shows the results for our SMT system when trained on the QCA parallel corpus, which was segmented using different training models of Morfessor with **B** = 40. The result for segmented Qatari Arabic is always better than the baseline, irrespective of the training model used for segmentation. We can see that the Morfessor model trained on a large monolingual corpus, i.e., on (ii) or (iii), yields better results.

Morfessor	**BLEU**	**OOV%**
Baseline	12.2	16.6
QCA	12.5	0.6
AVIA$_{QA}$, Novels	13.5	0.8
QCA, AVIA$_{QA}$, Novels	13.4	0.7

Table 3: Study of the effect of varying the training datasets for Morfessor on the Qatari to English SMT. "Baseline" shows the output of the MT system with no segmentation.

B	**10**	**40**	**70**	**100**	**130**
BLEU	13.3	13.5	**13.8**	12.9	12.6
OOV	0.3	0.8	1.4	2.8	2.8
After merging					
BLEU	12.5	13.4	**13.7**	12.8	12.3
OOV	1.5	1.9	3.9	6.5	9.8

Table 4: The effect of varying the perplexity threshold parameter *B* of Morfessor on SMT quality. "After merging" are the results using the post-processed Qatari segmented data.

The high reduction in OOV in Table 3 is because of the fine-grained segmentation. We tried different values for the perplexity parameter *B* in order to find a good balance between better BLEU scores and linguistically correct segmentations. The first part of Table 4 shows the effect of different values of *B* on the quality of the machine translation system trained on AVIA$_{QA}$, Qatari Novels. We achieved the best SMT score at $B = 70$.

We further analyzed the output of Morfessor at B = 70 and we noticed that it tends to generate very small segments of length two and three characters long. The segmentation produces more than one stem in a word and does not generate legal word units. For example, the word والصناعة 'and the industry' /wAlSinAEp/ is segmented as PRE/و + PRE/ال + STM/ص + PRE/ن + PRE/ا + STM/ع + SUF/ة. We apply a post-processing step that merges all stems in a word and affixes between them to one stem. So, a word can have only one stem. For example, the word والصناعة would be segmented as PRE/وال + STM/صناع + SUF/ة. This yielded linguistically correct segmentations in many cases. The second part of Table 4 shows the effect of the post-processing on the BLEU score. We can see that it remains almost the same with an increase in OOV rate.

For rest of the experiments in this paper, we used a value of 70 for the perplexity threshold parameter plus the post-processing on segmentation. We trained Morfessor on the concatenation of QCA, AVIA$_QA$ and Novels.[7]

Using other Arabic variations: In this section, we present experiments using MSA, EGY and LEV to English bitexts combined with the QCA bitext for Qatari Arabic to English machine translation. We explored three segmentation options for the Arabic side of the data: (i) no segmentation, (ii) ATB segmentation, and (iii) unsupervised segmentation using Morfessor.

The QCA corpus is of much smaller size compared to other Arabic variants, say MSA. It is possible that in the training of the machine translation models, the large corpus dominates the QCA corpus. In order to avoid that, we balanced the two corpora by replicating the smaller corpus X number of times in order to make it approximately equal to the large corpus (Nakov and Ng, 2009).[8] The complete procedure is described below.

In a nutshell, for building a machine translation system using the MSA plus Qatari corpus, we first balanced the Qatari corpus to make it approximately equal to MSA and concatenated them. For training Morfessor, the Qatari Arabic data consisted of QCA, Novels and AVIA$_{QA}$, while for SMT, it consisted of QCA only. In both cases, we balanced it to be approximately equal to MSA. We then trained Morfessor on the balanced (QCA, Novels, AVIA$_{QA}$) plus MSA data and we segmented the Arabic side of the balanced QCA plus MSA training data for machine translation. We built a machine translation system on the segmented data. We segmented the testing and tuning data sets similarly. We used the same balancing when we combined EGY-EN and LEV-EN with the Qatari Arabic – English data.

We also tried training multiple unsupervised models, but this yielded lower SMT quality compared to using a single model trained on multi-dialects. Using different models could result in having different segmentation schemes, which might not help in reducing the vocabulary mismatch between different variants of Arabic.

Train	NONE	ATB	Morfessor
QCA	12.2	12.9	**13.7**
'QCA,MSA	12.7	13.3	**14.6**
'QCA,EGY	13.0	13.5	**14.5**
'QCA,LEV	13.8	13.7	**15.2**

Table 5: BLEU scores for Qatari Arabic to English SMT using three different segmentation settings. 'QCA means the modified QCA corpus with number of tokens approximately equal to MSA, EGY and LEV in the respective experiments.

Table 5 shows the results. There are two things to point here. First, the SMT systems that used the unsupervised morphological segmenter, Morfessor, outperformed the systems that used no segmentation and those using the ATB segmentation. The Morfessor-based systems showed consistent improvements compared to the ATB-based systems over the no-segmentation systems. This validates our point that unsupervised morphological segmentation generalizes well for a variety of dialects and these SMT results complement that. The second observation is that adding a bitext for other dialects and MSA improves machine translation quality for Qatari–English SMT.

6 Translation into Qatari Arabic

Our monolingual corpora of Gulf subdialects could be also helpful when translating English into Qatari Arabic. We conducted a few basic experiments in this direction but without segmentation.

We trained an English to Qatari Arabic SMT system on the QCA bitext, using the same settings as described in Section 5. We then normalized the output of the translation system using the QCRI-Normalizer (Sajjad et al., 2013b).[9] As a language model, we used the Arabic side of the QCA corpus, novels and forum data, standalone and together. Table 6 presents the results of the effect of varying the language model on the quality of the SMT system. The best system shows an improvement of 0.22 BLEU points absolute compared to the baseline system that only uses the Arabic side of the QCA corpus for LM training.

The SMT system achieved the largest gain when adding QA forum data to the QCA data. SA and AE monolingual data also showed good improvements. This might be due to their relatively large sizes; we need further investigation.

[7] We did not see a big difference in training Morfessor with and without the QCA corpus, and we decided to use the complete data for training.

[8] Due to the spoken nature of the QCA corpus, it contains shorter sentences. Thus, we balanced the corpora based on the number of tokens rather than on the number of sentences.

[9] http://alt.qcri.org/tools/

LM	BLEU
QCA	2.78
QCA+QA-Novels	2.64
QCA+QA-Novels+BH-Novels	2.86
QCA+QA-Novels+KW-Novels	2.78
QCA+QA-Novels+AE-Novels	2.92
QCA+QA-Novels+SA-Novels	2.96
QCA+ALL-Novels	2.80
QCA+QA-Novels+QForum	**3.00**

Table 6: Results for English to Qatari SMT for varying language models. In all cases, the translation model is trained on the QCA bitext only.

Note the quite low BLEU scores, especially compared to the reverse translation direction. One reason is the morphologically rich nature of Qatari Arabic, which makes translating into it a hard problem. The small amount of training data further adds to it. We expect to see larger gains compared to Qatari Arabic to English machine translation when segmentation is used.

7 Conclusion and Future Work

We have demonstrated the feasibility of using an unsupervised morphological segmenter to increase the resource adaptability of Arabic variants. We evaluated the segmentation on a Qatari dialect by building a Qatari Arabic to English machine translation system. We further adapted MSA, EGY and LEV in the simplest machine translation settings and we showed a consistent improvement of 1.5 BLEU points when compared to the respective baseline system that uses no segmentation.

In the future, we would like to explore the impact of segmentation on both the translation model and the language model when translating into Qatari Arabic. This involves greater challenges, as a desegmenter is required for the translation output with every segmentation scheme.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May.

Rania Al-Sabbagh and Roxana Girju. 2010. Mining the web for the induction of a dialectical Arabic lexicon. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta, May.

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *Proceedings of the 6th International Conference on Informatics and Systems*, Cairo, Egypt, March.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), June.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, Trento, Italy, May.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), January.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR, June.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a TV broadcasts speech recognition system for Qatari Arabic. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May.

Spence Green and John DeNero. 2012. A class-based agreement model for generating accurately inflected translations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, July.

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 the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, July.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, April.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for Egyptian Arabic. In *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology*, Montreal, Canada, June.

Nizar Y Habash. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1), August.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, UK, July.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Scotland, UK, July.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic. In *Proceedings of the Arabic Natural Language Processing Workshop*, Doha, Qatar, October.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for ngram langauge modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, Michigan, May.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, Edmonton, Canada, May.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demonstration Program*, Prague, Czech Republic, June.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, MA, May.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, Italy, May.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the Arabic Natural Language Processing Workshop*, Doha, Qatar, October.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Suntec, Singapore, August.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for sentence-level BLEU+1 yields short translations. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, December.

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 on association for computational linguistics*, Philadelphia, PA, July.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Denver, CO, June.

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*, MA, USA, August.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013a. Translating dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August.

Hassan Sajjad, Francisco Guzman, Preslav Nakov, Ahmed Abdelali, Kenton Murray, Fahad Al Obaidli, and Stephan Vogel. 2013b. QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic Spoken Language Translation. In *Proceedings of the 10th International Workshop on Spoken Language Translation*, Hiedelberg, Germany, December.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the youtube dialectal Arabic commentary corpus. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May.

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*, Edinburgh, Scotland, July.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*, Denver, CO, October.

Vesa Siivola, Mathias Creutz, and Mikko Kurimo. 2007. Morfessor and VariKN machine learning tools for speech and language technology. In *Proceedings of the 8th International Conference on Speech Communication and Technology (Interspeech)*, Antwerpen, Belgium, August.

Ming-Feng Tsai, Preslav Nakov, and Hwee Tou Ng. 2010. Morphological analysis for resource-poor machine translation. Technical report, Kent Ridge, Singapore, December.

Omar F Zaidan and Chris Callison-Burch. 2011. 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: short papers-Volume 2*, Portland, OR, June.

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*, Montreal, Canada, June.

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. 2006. Bridging the inflection morphology gap for Arabic statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York, NY, June.