Morphology-Aware Word-Segmentation in Dialectal Arabic Adaptation of Neural Machine Translation

Ahmed Y. Tawfik, Mahitab Emam, Khaled Essam
Robert Nabil and **Hany Hassan**
Microsoft
atawfik|a-maemam|a-kessa|a-ronabi|hanyh@microsoft.com

Abstract

Parallel corpora available for building machine translation (MT) models for dialectal Arabic (DA) are rather limited. The scarcity of resources has prompted the use of Modern Standard Arabic (MSA) abundant resources to complement the limited dialectal resource. However, clitics often differ between MSA and DA. This paper compares morphology-aware DA word segmentation to other word segmentation approaches like Byte Pair Encoding (BPE) and Sub-word Regularization (SR). A set of experiments conducted on Egyptian Arabic (EA), Levantine Arabic (LA), and Gulf Arabic (GA) show that a sufficiently accurate morphology-aware segmentation used in conjunction with BPE or SR outperforms the other word segmentation approaches.

1 Introduction

Building machine translation models for resource constrained languages can benefit from parallel corpora available in related languages. Vocabulary adaptation (Passban et al., 2017) has been used to train statistical and neural machine translation models for Azeri, a resource constrained language, leveraging its similarity to Turkish. Projection to a universal representation language (Gu et al., 2018) generates high quality machine translation model for a resource constrained language given a set of related resource-rich languages.

Research in dialectical Arabic translation tried to leverage the resources available in Modern Standard Arabic (MSA) using several techniques. Starting with statistical and rule-based methods for transforming DA to MSA (Al-Gaphari and Al-Yadoumi, 2012), and evolving to generating DA data from MSA parallel data using semantic projections (Hassan et al., 2017), and multi-task learning of part-of-speech tagging and machine translation to guide the translation model towards leveraging the grammatical roles in translation (Baniata et al., 2018). While earlier statistical and rule-based cross-dialectical techniques managed to leverage morphological word segmentation, more recent attempts have largely abandoned morphological segmentation in favor of language agnostic segmentation techniques like Byte Pair Encoding (BPE) (Sennrich et al., 2016) and Sub-word Regularization (SR) (Kudo, 2018). In fact, these learned language agnostic word segmentation have proved that they can rival morphological segmentation in neural MT. In a translation task from language D to language E, if language D (say an Arabic dialect) and language A (say modern standard Arabic) are two closely related languages such that a word W_A in language A is semantically equivalent to a word W_D in language D. Moreover, we assume that these two words share a common stem but have different clitics. So, the two words can be morphologically segmented as follows: $W_A = P_A R S_A$, and $W_D = P_D R S_D$ where P_A is a sequence of zero or more characters forming the prefix of W_A. Similarly, S_A is a sequence of characters forming the suffix of W_A, while P_D and S_D denote the prefix and suffix of W_D, and R is the shared root or stem.

Due to the limited training data for the language pair $\{D, E\}$, the root R is one that we hope to learn from the abundant data for the pair $\{A, E\}$. Intuitively, a morphology-aware word segmentation is more likely to produce the correct prefixes and suffixes, making it easier to learn the translation of R to E. As clitics tend to occur frequently, the MT model would have learned their translation from the scarce resources for the pair $\{D, E\}$; thus, successfully translation an out-of-vocabulary word for the $\{D, E\}$ pair. For illustration consider the example in Table1 below. The dialectal Egyptian word "هيقولوا" [hayqwlwA] is segmented into four segments. Similarly, the correspond-

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 11–17
Florence, Italy, August 1, 2019.

Segmented Dialectal Word	هـ#يـ#قول#وا ha#y#qwl#wA
Segmented MSA Word	سـ#يـ#قول#ون sa#ya#qwl#wn
English Translation	They will say
Alignment MSA-EN	0-1;2-2; 1,3-0
Alignment DA-EN	0-1;2-2; 1,3-0

Table 1: Illustrative word segmentation example

ing MSA word "سيقولون"[sayaqwlwn]. Both words share the same stem "قول" [qwl], that can be learned from the resource rich MSA, while the dialectal future marking dialectal prefix "هـ#" [ha#] can be learned from other future tense verb in the training data. Similarly, the plural 3rd person markers can be learned from other verbs in the resource constrained parallel data. The alignments in the table are zero based word index alignment from Arabic to English.

The question that this paper aims to address is whether morphological word segmentation still has an advantage over language agnostic methods, in the context of leveraging parallel data in a resource-rich language to improve the MT of a related resource constrained one. This question is particularly interesting when we consider morphologically-rich languages like Arabic and its dialects. The remainder of this paper introduces the role of word segmentation in machine translation in Section 2. This section also reviews popular word segmentation techniques and introduces the morphology-aware segmentation approach that is used in our experiments. Section 3 reviews the neural machine translation approach that we use to train and adapt translation models for dialectal Arabic. Section 4 presents the experiments that we conducted along with their results. Section 5 reviews some related works. Finally, Section 6 summarizes the findings and concludes the paper.

2 Word Segmentation in NMT

The size of vocabulary found in a typical English dictionary is less than 100,000 words. A vocab around 16,000 words, provides 98% coverage for the Brown corpus. However, due to its agglutinative nature, the size necessary to achieve similar coverage for Arabic, whether standard or dialectal, is much larger. The size of the vocabulary extracted from the Arabic Gigaword corpus (Parker et al., 2009) exceeds 800,000 words.

Such vocab sizes are well beyond what current technology can handle efficiently. Therefore, it is common to use word segmentation for highly agglutinative languages like Arabic, or highly compounding languages like German (Huck et al., 2017), and more generally, for any large vocab NMT system. Two popular language agnostic word segmentation techniques are Byte-Pair-Encoding (BPE) (Sennrich et al., 2016) and Subword Regularization (SR) (Kudo, 2018).

2.1 Byte-Pair-Encoding (BPE)

Originally conceived as compression algorithm (Gage, 1994), BPE is a greedy technique often used to segment words into common subwords as a preprocessing step in a NMT training pipeline (Sennrich et al., 2016). BPE starts by splitting all the words in the training lexicon into individual characters, and proceeds by merging frequent character sequences until reaching a specified number of merge operations. Thus, by the end of the algorithm most frequent word segments would have been joined into a single symbol. The resulting trained segmenter is stored and applied to test and runtime inputs.

2.2 Subword Regularization (SR)

Subword Regularization (Kudo, 2018) generates probabilistic word segmentations to make the NMT training more robust. The probabilities of the segments are computed from a unigram language model defined over subword symbols. The intuition behind it is that if a sentence is represented by using multiple subword sequences it will produce some regularization during the training thus making the machine translation model more robust. The results achieved using SR, depends on the setting of three parameters: the vocab size, the size of n-best segmentation, and a smoothing parameter that controls the probabilistic sampling of segmentation.

2.3 Linguistically Motivated Segmenter

The problem with BPE and Subword Regularization is that they don't take into consideration any information about the language which might cause a loss of semantic and syntactic properties such as inflection and composition. These syntactic features are potentially useful in machine translation as semantic modifiers. The importance of using a linguistically motivated segmenter has been shown previously (Huck et al.,

2017) as they assist greatly in reduction of vocabulary size while helping improve the translation of unseen words (open vocabulary translation problem). The linguistically aware dialectal Arabic segmenter used in this work is a retrained version of the Unified Dialectal Arabic Segmenter (UDAS) (Samih et al., 2017). The unified segmentation model is based on a bidirectional Long Short-Term Memory (bi-LSTM) Recurrent Neural Network (RNN) that is coupled with Conditional Random Fields (CRF) sequence labeler trained to segment words from four different dialects namely Egyptian (EGY), Levantine (LEV), Gulf (GLF), and Maghrebi (MGR). The segmenter leverages the observation that different Arabic dialects do not only share vocabulary and some morphological properties with MSA, but they also share some commonalities amongst each other. Thus, a single model provides higher accuracy than a dialect specific model while eliminating the need for dialect identification before segmentation. This segmenter operates directly on raw text without requiring any preprocessing or word normalization while employing a lookup scheme that use segmentations that are seen in training directly during testing in order to improve the performance and the accuracy of segmenting a words into prefixes, stems and suffixes. To improve the segmentation model, we added to the training data, publicly available data from the LDC-Arabic Treebank (LDC2010T08, LDC2010T13, and LDC 2011T09), as well as dialectal Arabic treebanks (LDC2016T02, LDC2016T18, and LDC2018T23) to reach a total of 231,846 segmented sentences. Table 2 presents the accuracies of the segmentation for each dialect compared to the accuracy in the baseline model (Samih et al., 2017). To measure the accuracy, a 20% subset of the original UDAS training data is set aside as unseen testset. Despite some inconsitencies in segment labeling in the various datasets, the addition of data has resulted in improvements for all dialects. Like the original UDAS model, a lookup table has proved helpful in improving the trained model. We populated the lookup table with words found in the training data that the trained model fails to segment. The accuracy improvements were slightly higher for Egyptian which can be attributed to the fact that the added data had a large portion in that dialect (LDC2018T23).

	EGY	GLF	LEV	MGR
Retrained Model	99.4	98.9	96.2	96.1
Baseline	95.3	93.1	93.9	91.4

Table 2: Accuracy of the retrained unified dialectal segmenter compared to the baseline model (Samih et al., 2017).

3 NMT Training for Dialectal Arabic

To train Neural Machine Translation (NMT) for Arabic dialects, we use the now ubiquitous encoder-decoder structure. In these structures, the encoder maps a source language input to a dense internal vector representation, that the decoder maps to a corresponding target language output. Like other languages, a recurrent neural network (RNN-based) with attention (Bahdanau et al., 2015) or a feed-forward network with multi-attention (Transformer-based) (Vaswani et al., 2017), Sequence to Sequence architectures are used for the encoder and the decoder. Dialectical Arabic parallel resources are very scarce compared to the amount of data necessary to train general purpose NMT models. The parallel data publicly available for Arabic dialects used in this work are limited to:

- Crowd sourced translations for Levantine and Egyptian (LDC2012T09, (Zbib et al., 2012)),
- BOLT Egyptian Arabic parallel discussion forums data (LDC2019T01),
- Qatari Arabic Corpus that includes English translation for several hours of Qatari TV broadcast conversations.
- Dialectal contents extracted from the Arabic subtitles using a dialect ID trained fastText language ID type model (Joulin et al., 2017).
- Translation of the Egyptian Callhome (Kumar et al., 2014) a crowd-sourced translation of a conversational telephony dataset.

The total number of parallel sentences for each dialect ranges from tens of thousands for gulf to several hundreds of thousands for Egyptian and Levantine. These amounts are well below the minimum required for an adequate coverage for a language. Therefore, to leverage the abundant MSA resources, we train a base model using MSA data along with the limited amount of dialectal data. We use domain adaptation techniques to fine tune

Dialect	Training Set	DevTest
MSA	2.5 M sent.	-
Gulf	38 K sent.	2 K sent.
Levantine	219 K sent.	2 K sent.
Egyptian	502 K sent.	2 K Callhome.

Table 3: Training and test corpora sizes

the base model (Freitag and Al-Onaizan, 2016). Our methodology is different in that whereas they train the model on the out-of-domain data only and then adapt to the in-domain data, we train on the joint data to allow the model to learn dialect-specific vocab and then adapt using the in-domain data. Also, whereas they use an ensemble of the base model and the adapted model, we use an ensemble of two adapted models. Arabic dialects have some common words and idioms which overlap with MSA. So, when training a dialectical models it's beneficial to first train the model with the high-resourced MSA data jointly with the dialectical data with optional duplication so that the dialectical vocab is significant in the training data and doesn't get pruned or overwhelmed by the MSA vocab, and then adapting the model by training it for a few epochs with a small learning rate on the relatively small dialectical data to bias the model further to the dialect in the cases where the meaning in the dialect is different from the meaning in MSA.

4 Experiments and Results

Several experiments were conducted to examine the impact of the dialectical segmenter on the quality of the MT system built with it for a resource constrained languages, and how it compares to other segmentation techniques like BPE and SR. The experiments were carried out using Marian v1.7.6 (Junczys-Dowmunt et al., 2018) a public neural machine translation framework which supports sentence piece tokenization with its two variant BPE and SR (unigram language model) as well as word tokenization which is basically tokenizing the corpus on white spaces. Most parameters of Marian were the same as the defaults except for the validation set settings which were adapted to each dialect according to the size of its data.

Table 3 shows the distribution of the training and test data sizes used in the experiments. For the Gulf and Levantine dialects, 2000 sentences are set aside and equally divided into validation and test. For Egyptian, the callhome validation and test split is used after disfluency removal. The disfluency removal consists of removing incomplete words, filler words, and repeated words. This processing is necessary because we started with the speech transcripts (LDC97T19, LDC2002T38) which have full verbatim transcripts of the corresponding speech corpora. As described in Section 3, the base model training merges Arabic dialect sentences and MSA. Therefore, special care was needed to train the MT system for the Gulf dialect because it has far fewer sentences than MSA we needed to duplicate the Gulf data 10 times in order to make the sizes of the data of the Gulf dialect and other dialects comparable. The adaptation uses the dialect data only to fine-tune the base model trained for that dialect at a lower learning rate.

Gulf – English Results		
Word Segmentation	**Base BLEU**	**Adapt BLEU**
Dialectal segmenter	12.36	12.64
BPE	13.19	13.36
SR	14.08	14.30
Dialectal segmenter + BPE	**14.58**	**14.58**
Dialectal segmenter + SR	14.18	14.18
Levantine – English Results		
Word Segmentation	**Base BLEU**	**Adapt BLEU**
Dialectal segmenter	19.41	19.98
BPE	20.83	21.56
SR	20.42	21.81
Dialectal segmenter + BPE	21.9	22.47
Dialectal segmenter + SR	**22.07**	**23.08**
Egyptian – English Results		
Word Segmentation	**Base BLEU**	**Adapt BLEU**
Dialectal segmenter	37.22	37.86
BPE	36.19	36.83
SR	36.79	37.76
Dialectal segmenter + BPE	36.93	**38.2**
Dialectal segmenter + SR	**37.44**	37.68

Table 4: The word segmentation technique, base model BLEU score, and adapted model BLEU score for each of the three dialects.

As summarized in Table 4, for each dialect, we evaluated five word segmentation approaches:

1. The dialectal segmenter as the only segmenter.

2. Byte-Pair Encoding (BPE) as the only segmenter.

3. Subword Regularization (SR) as the segmenter.

4. Byte-Pair Encoding applied to dialectically segmented corpora.

5. Subword Regularization applied to dialectically segmented copora.

In all cases, the vocab was kept at 40 K subwords. For the base models in all three dialects, the best performing word segmentation combined dialectal segmentation with either BPE or SR. This continued to be the case after adaptation. The gain attributable to dialectal segmentation[1] was 0.28 BLEU point for Gulf, 1.27 for Levantine, and 0.44 for Egyptian. It also worth noting that Subword Regularization has consistently outperformed BPE alone. The low scores for the Gulf dialect are due to the small size of the test set and the use of a highly dialectal spelling in the data that limited the model's ability to benefit from the MSA training. While Levantine and Egyptian training data are comparable in size, the BLEU scores reported for Egyptian are based on 4 reference translations, while Levantine scores use a single reference. To assess the similarity of word segmentation obtained by the various approach, we computed the Levenshtein edit distances between the segmented sentences for a random subset of 150 dialectal Arabic sentences. In this set, no two segmentation techniques produced the same word segmentation for all the words in any sentence. However, applying SR or BPE to a dialectically segmented sentences gives very similar segmentations with an average edit distance of 2.55 per sentence. The segmentations obtained by BPE and SR were also relatively similar with an average edit distance of 5.03.

Table 5 summarizes the average number of edits necessary to map a segmented sentence using one approach to the others. In the table, DS is the dialectal segmenter. The relatively large number of edits between the dialectal segmenter and both BPE and SR suggest that these language agnostic approaches have not fully captured the morphological aspects of Arabic dialects.

[1] Calculated as BLEU difference between the best adapted model with dialectal segmentation and the best adapted model without dialectal segmentation

5 Related Research

Translating Arabic dialects has been a focus with the machine translation community. In statistical machine translation (SMT), the use of morphology-aware word segmentation for Arabic has been studied (Lee et al., 2003),and (Habash, 2007). Sajjad et al. 2013 maps DA closer to MSA prior to translation. Sawaf 2010 also uses dialect normalizations and uses morphological for the dialects as well as MSA. This technique has significantly reduced the vocabulary size. However, the new vocab size restriction imposed by NMT and the advent of newer language independent word segmentation techniques like BPE and SR, as well as the advances in dialectal Arabic word segmentation prompted us to revisit the topic. Within the NMT context, Huck et al. 2017 studied the impact of linguistically-aware word segmentation on the translation from English to German. In their work, the linguistically aware techniques show some gains from combining linguistically-aware segmentation with BPE. In our work, we have observed similar gains from the combination with BPE, which suggests that such gains may be reproducible for other morphologically complex languages.

6 Conclusions and Future Work

Learning dialectal segmentation using a unified model (Samih et al., 2017) for the various dialects can achieve high accuracies provided sufficient training data. In our experiments, a segmentation accuracy of 99.4% was reached for Egyptian Arabic. Significant improvements were also achieved for other dialects. Our hypothesis has been that a high accuracy dialectal segmenter would maximize the transfer between the resource rich MSA machine translation and the resource restricted Arabic dialects. The experimental results seem to confirm that there is some advantage from using a high accuracy dialectal segmenter jointly with a language independent word segmentation technique like Byte-Pair Encoding or Subword Regularization. However, in using Subword Regularization in our experiments, we relied on the default values for the n-best size and smoothing as implemented in Marian. It would be interesting to see if our observations will continue to hold if these parameters are carefully tuned.

	BPE Only	BPE + DS	SR + DS	SR only
DS Only	11.47	10.91	11.51	9.71
BPE Only		15.10	16.37	5.03
BPE + DS			2.55	14.64
SR + DS				14.10

Table 5: Average Lenvenshtein Edit Distance between segmented sentences

References

Ghaleb Al-Gaphari and Mohammed Al-Yadoumi. 2012. A method to convert Sana'ani accent to modern standard arabic. *International Journal of Information Science and Management (IJISM)*, 8(1):39–49.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *Proc. International Conference on Learning Representations*.

Laith H. Baniata, Seyoung Park, and Seong-Bae Park. 2018. A multitask-based neural machine translation model with part-of-speech tags integration for arabic dialects. *Applied Sciences*, 8(12):2502.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. *Proceedings of the 11th MT Summit*, 10.

Hany Hassan, Mostafa Elaraby, and Ahmed Y. Tawfik. 2017. Synthetic spoken data for neural machine translation. In *IWSLT*, pages 82–89.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *Proc. of ACL*.

Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. 2014. Translations of the callhome Egyptian Arabic corpus for conversational speech translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*. Citeseer.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language model based Arabic word segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 399–406. Association for Computational Linguistics.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2009. Arabic gigaword.

Peyman Passban, Qun Liu, and Andy Way. 2017. Translating low-resource languages by vocabulary adaptation from close counterparts. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):29.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal arabic to english. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 1–6.

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, and Laura Kallmeyer. 2017. Learning from relatives: unified dialectal Arabic segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 432–441.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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. Association for Computational Linguistics.