Multi-Lingual Phrase-Based Statistical Machine Translation for Arabic-English

Ahmed Bastawisy and **Mohamed Elmahdy**
Computer Science Department
German University in Cairo, Cairo, Egypt
ahmed.bastawisy@student.guc.edu.eg, mohamed.elmahdy@guc.edu.eg

Abstract

In this paper, we implement a multi-lingual Statistical Machine Translation (SMT) system for Arabic-English Translation. Arabic Text can be categorized into standard and dialectal Arabic. These two forms of Arabic differ significantly. Different mono-lingual and multi-lingual hybrid SMT approaches are compared. Mono-lingual systems do always result in better translation accuracy in one Arabic form and poor accuracy in the other. Multi-lingual SMT models that are trained with pooled parallel MSA/dialectal data result in better accuracy. However, since the available parallel MSA data are much larger compared to dialectal data, multi-lingual models are biased to MSA. We propose in the work, a multi-lingual combination of different mono-lingual systems using an Arabic form classifier. The outcome of the classier directs the system to use the appropriate mono-lingual models (standard, dialectal, or mixture). Testing the different SMT systems shows that the proposed classifier-based SMT system outperforms mono-lingual and data-pooled multi-lingual systems.

1 Introduction

The Arabic language is the largest still living Semitic language. Arabic is spoken by more than 350 million people around the world. It is also one of the five official languages of the United Nations, and the first official language of twenty-two countries known by the Arab world. Arabic is also used as a second language for more than 1.2 billion people.

Modern Standard Arabic (MSA) is currently considered the formal Arabic variety across all Arabic people. MSA is used in news broadcasts, newspapers, formal speech, books, movies subtitling, and whenever the target audience or readers come from different nationalities. However, MSA is not the natural language for everyday life communications and on social networks. In fact, dialectal Arabic is usually used in this case.

A major problem in all Arabic Natural Language Processing tasks, and in particular Statistical Machine Translation (SMT) is the existence of the Arabic dialects. There exist significant syntactic, morphological, and lexical differences between MSA and the different Arabic dialects. That is why they are sometimes considered as completely different languages (Soudi et al., 2012; Elmahdy et al., 2012)

There were big efforts exerted to improve Arabic-English SMT, most of these efforts were focused on MSA rather than dialectal Arabic. This is mainly due to the fact that the vast majority of available parallel Arabic data are for MSA, whilst relatively sparse and limited parallel data are available for dialectal Arabic (Alqudsi et al., 2014).

To tackle the problem of dialectal Arabic parallel data sparsity, in many previous, they have normalized dialectal words/phrases into corresponding MSA equivalents. This normalization, or pivoting, is basically a rule-based approach to paraphrase dialectal words into MSA. This normalization would allow the usage of existing MSA SMT systems (Salloum and Habash, 2013; Sawaf, 2010).

In (Zbib et al., 2012), instead of relying on normalization or pivoting, they have collected extra dialectal Arabic parallel in combination to existing MSA data. Results showed that the proposed pooling technique has improved translation accuracy for dialectal Arabic. However, MSA translation accuracy has slightly decreased.

Because of the complex morphological nature of Arabic, some prior work, as in (Lee, 2004), fo-

Proceedings of Recent Advances in Natural Language Processing, pages 86–89,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_013

cused on MSA morphological analysis to improve Arabic SMT.

The aim of this work is to build a Multilingual Arabic SMT system that supports MSA as well as dialectal Arabic. Another goal is that the addition of dialectal Arabic should not affect MSA translation accuracy. Moreover, since available MSA data are always larger than dialectal data, the system should not be biased to MSA.

In this paper, we propose training three different Arabic SMT models. One model for MSA, another system for Dialectal Arabic, and the last one is a hybrid model that is trained with a data pool of parallel Arabic-English for MSA and dialectal Arabic. A pre-classifier is built to choose the appropriate model to be used.

2 Translation Models

Throughout this work, all translation models were built using Giza Aligner and Moses SMT engine (Philipp et al., 2007). Three translation models have been created: MSA-English model, dialectal-English model, and hybrid-English model. To train the MSA-English translation model, a parallel dataset of 26M words was utilized from the ISI Arabic-English Automatically Extracted Parallel Text corpus (Dragos and Daniel, 2007). An independent MSA-English evaluation set of 300K words was used to tune the model. A MSA-English test set of 300K words is used to evaluate MSA-English translation accuracy.

To train the dialectal-English translation model, a parallel dataset of 2.7M words was utilized from the Arabic-Dialect/English Parallel Text corpus (Technologies et al., 2012) (notice the huge difference between the size of available MSA and dialectal data). An independent dialectal-English evaluation set of 300K words was used to tune the model. A dialectal-English test set of 300K words is used to evaluate MSA-English translation accuracy.

The hybrid translation model has been trained by pooling both training sets of MSA and dialectal parallel data that consists of 26M MSA words and 2.7M dialectal words. Model tuning was performed using the two evaluation sets of MSA and dialectal Arabic.

A statistical tri-gram language model is trained for English. Language model training set consists of 688M words from 2011 and 2012 articles (News Crawl) that is described in (Sofia, 2013). The English language model is used to estimate the prior probability in all of the proposed SMT techniques.

The three translation models have been tested with the three testing sets (MSA, dialectal, MSA+dialectal). As shown in Table 1, the MSA model has resulted in BLEU score of 34.8, 2.6, and 18.7 on MSA, dialectal, and MSA+dialectal testing sets. It is clear that the MSA model performs poorly on dialectal Arabic data. Using dialectal Arabic model, the results were 4.1, 15.9, and 10.0 on MSA, dialectal, and MSA+dialectal respectively. It is clear that the dialectal model performs better on dialectal data, and performs poorly with MSA data. The hybrid model has resulted in a better acceptable accuracy across both MSA and dialectal Arabic. The hybrid model has resulted in 33.2, 12.3, and 22.8 BLEU for MSA, dialectal, and MSA+dialectal respectively. The hybrid model seemed to be a little bit biased towards MSA as the relative decrease in the accuracy was -4.6% relative the MSA baseline model, and -22.6% relative to the dialectal baseline model.

Translation model	Parallel data type		
	MSA	Dialect.	MSA+Dialect.
MSA	34.8	2.6	18.7
Dialectal	4.1	15.9	10.0
Hybrid	33.2	12.3	22.8

Table 1: BLEU score for the different SMT systems on MSA, dialectal, and MSA+dialectal data.

3 Classification-Based Translation

Although before adding the classifier, MSA and Dialectal Arabic-English SMT systems accuracy were poor across the different variants, the hybrid system that was trained with both MSA and dialectal data has resulted in better accuracy. However, the aim of the Classification-Based Translation is to further improve the accuracy across both dialectal and MSA, and to overcome the bias problem of the hybrid model.

Two classification techniques have been used, the first technique is to classify input Arabic text into two classes Standard and Dialectal, and accordingly translate them with the appropriate system. The second technique is to classify input Arabic text into three classes Standard, Hybrid and Dialectal, and then use the appropriate system ac-

cordingly.

A tri-gram MSA language model is built for the sake of classification. More than 355M words from the Arabic Gigaword corpus (Parker et al., 2011) were used to train a MSA language model.

The MSA language model is used in text classification by scoring every input sentence by the language model. Sentences with high log likelihood are classified as MSA, whilst sentences with low log likelihood are classified as Dialectal.

3.1 First Classification Techniques

In the techniques, text segments are classified into two categories: MSA or dialectal. Two-passes optimization search was made to find the optimal language model scoring threshold between MSA and dialectal classes.

In the first pass, a coarse search was performed by varying classification threshold from 0.0 to -10.0 with a coarse step of 1.0. For each iteration, classification accuracy is evaluated. The initial optimal threshold was found to be -4.0 which has resulted in classification accuracy of 95.58% on the evaluation sets of MSA and dialectal Arabic.

In the second optimization pass, a fine step search was performed around the initial -4.0 threshold with a variable value of -3.0 to -5.0 with a step of 0.1. Figure 1 shows classifier's accuracy test with a fine step of 0.1 ($x = x - 0.1$). As shown in the graph, the optimal threshold is -3.7 which has resulted in classification accuracy of 96.64%. Thus, threshold of -3.7 has been used.

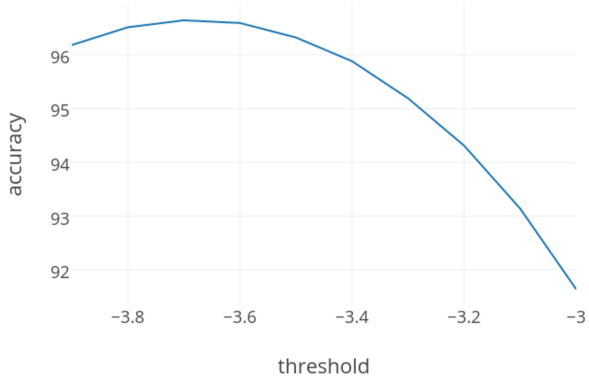

Figure 1: Fine tuning graph for MSA/dialectal classification threshold.

In this technique, the classifier works on classifying the test set and generating two file groups, the first group contains the MSA classified segments which have scored more than the threshold -3.7. The second group contains the dialectal Arabic classified segments which have scored below or equal the score threshold -3.7.

After that, each classified Arabic text file was translated with the corresponding SMT system, and then all translations were evaluated with the BLEU score test.

3.2 Second Classification Technique

In this technique, instead of having a sharp threshold between MSA and dialectal classes, we have created a window with the optimal threshold in the middle. Any sentence with a score that lies in this window is classified with a third class. That class is labeled the *mixture class*. It is assumed that any sentence in this class (very close to the threshold) might contain a mixture of dialectal and MSA words, which is a common case on social media for instance. The optimal window range has been found to be from -2.7 to -5.45. The three classes in this case are: Dialectal, MSA, and mixture.

The test set is classified into three file groups, the first group contains MSA sentences, which has scored more than the window upper bound -2.7, the second group has the Hybrid Arabic classified sentences, which has score within the window from -2.7 to -5.45, the third group has the Dialectal classified sentences, which has scored less than the window lower bound -5.45.

After that, each classified Arabic text file was translated with the corresponding SMT system, and then all translations were evaluated with the BLEU score test.

4 Experimental Results

The two classification-based translation techniques have been tested on a test set that combines both testing sets of MSA (300K words) and dialectal Arabic (300K words).

The first classification technique has resulted in a BLEU translation accuracy of 29.1 absolute outperforming the hybrid model with a relative increase in the accuracy of 27.6% as shown in Table 2.

The second classification technique has resulted in a BLEU translation accuracy of 29.0 absolute outperforming the hybrid model with a relative increase of 27.2%.

As shown in Table 2, both techniques have significantly improved translation accuracy in com-

parison to all of the three baseline systems. This means that introducing a pre-classification stage might be a helpful step in improving the performance of Arabic machine translation systems.

The BLEU score is slightly better in the first classification technique than the second one with an absolute difference of 0.1. This implies that it is enough to classify input Arabic text into just two categories instead of three.

Technique	BLEU	Relative
Hybrid Model	22.8	baseline
Classifier-based 1	29.1	+27.6%
Classifier-based 2	29.0	+27.2%

Table 2: Translation accuracy on MSA+dialectal parallel data for the hybrid model, classifier-based technique 1, and classifier-based technique 2.

5 Conclusions

This paper has focused mainly on enhancing the accuracy of SMT across MSA and dialectal Arabic. Three baseline Arabic-English SMT systems were built: MSA, dialectal, and Hybrid. MSA system resulted in significantly low accuracy on dialectal data, whilst dialectal system resulted in low accuracy on MSA data. The hybrid system performed with a better average accuracy across both MSA and dialectal data.

In order to classify input text into the correct variety of Arabic (dialectal or MSA), two classification techniques have been proposed. The first technique classifies the testing data into two categories, one to be translated with the MSA model, and the other to be translated with the dialectal model. The second technique classifies the testing data into three classes, one to be translated with the MSA model, one to be translated with the hybrid model, and the last one to be translated with the dialectal model.

Both techniques have significantly improved translation accuracy on a balanced testing set that contains equal amounts of MSA and dialectal data. The first technique resulted in a slightly better BLEU score than the second classification one.

References

Arwa Alqudsi, Nazlia Omar, and Khalid Shaker. 2014. Arabic machine translation: a survey. *Artificial Intelligence Review* 42(4):549–572.

Stefan Munteanu Dragos and Marcu Daniel. 2007. ISI Arabic-English Automatically Extracted Parallel Text LDC2007T08. Web Download. Philadelphia: Linguistic Data Consortium.

Mohamed Elmahdy, Rainer Gruhn, and Wolfgang Minker. 2012. *Novel Techniques for Dialectal Arabic Speech Recognition*. Springer-Verlag New York, 1 edition.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics, pages 57–60.

Parker, Robert, et al. 2011. Arabic Gigaword Fifth Edition (LDC2011T11). Linguistic Data Consortium.

Koehn Philipp, Hoang Hieu, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL).

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic. In *HLT-NAACL*. pages 348–358.

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*.

Sofia. 2013. News Crawl (articles from 2011 and 2012). web Download. Shared Task: Machine Translation.

Abdelhadi Soudi, Ali Farghaly, Gunter Neumann, and Rabih Zbib. 2012. *Challenges for Arabic Machine Translation*. Natural Language Processing 9. Benjamins, John.

Raytheon BBN Technologies, Linguistic Data Consortium, and Sakhr Software. 2012. Arabic-Dialect/English Parallel Text (LDC2012T09). Linguistic Data Consortium.

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*. Association for Computational Linguistics, pages 49–59.