Machine Translation for Arabic Dialects

Bashar Alhafni, Utkarsh Gera

Performing Machine Translation on Social Media data is a critical task for analyzing this data on various levels. When the goal is translating a particular source language to a target language, the availability and the format of the resources of these languages play an essential role in Machine Translation. Since most of the modern Machine Translation tasks are based on neural models, having many resources to train the models to translate a particular target language is essential. Therefore, with the current advances in neural models, it is really easy to translate high resource languages like English and Chinese, however, it’s not the case for low resources languages like Uzbek and Urdu. However, some languages are considered to have high resources in their standard form, but the different dialects of these languages are considered to have low resources, such as Arabic.

As the largest member of the Semitic languages, Arabic has more than 22 dialects and it’s spoken by more than 350 million people worldwide. There have been many attempts to apply Machine Translation to Arabic Social Media data but only a few of these attempts were successful because of low availability of these resources. One of the first attempts to translate Arabic dialects to English, was to normalize the various dialects to MSA (Modern Standard Arabic) and then perform the translation on the Modern Standard Arabic. By doing so, the task becomes much easier since the Modern Standard Arabic is considered to have high resources and the translation could be easily done by a well-trained neural model.

In their paper, *Morphological Analysis and Disambiguation for Dialectal Arabic,* Habash et al. describe the enhancements of a tool called MADA (Morphological Analysis and Disambiguation of Arabic) that is capable of MSA morphological analysis, disambiguation, POS tagging, tokenization, lemmatization and diacritization, to work for ARZ (Egyptian Arabic) the most commonly spoken Egyptian dialect. The enhancements enabled Habash et. al to outperform the older version of MADA that was directly applied on the Egyptian dialects for morphological tagging. These enhancements included a replacement of the original analyzer in MADA with an ARZ specific analyzer. Which solves the problem of mapping ARZ words into CODA (Conventional Orthography for Dialectal Arabic) while accepting a wide range of spelling variants. After the development of the new MADA tool, Sawaf et al. attempted to normalize the different Arabic dialects to MSA to perform Machine Translation to English and they were able to achieve about 1 BLEU (Bilingual Evaluation Understudy) increase on broadcast news and about 2 points on Web text.

However, in their paper, *Machine Translation for Arabic Dialects,* Zabib et. al proposed a novel approach to translate different dialects for Arabic to English and were able to accomplish a BLEU (Bilingual Evaluation Understudy) of 6.3 to 7.0 higher than the Modern Standard Arabic Machine Translation system. Their approach was to combine the data for various dialects and develop a parallel corpus for Levantine (one of the Arabic dialects) to English as well as Egyptian to English by using Amazon’s Mechanical Turk crowdsourcing services and other reliable translation services. After constructing their corpus, Zabib et. al performed a set of experiments to contrast systems trained by their corpus with systems trained on **much larger** MSA-English corpus. All the experiments use the same methods for training, decoding and parameter tuning and as expected, the systems trained with the various dialects to English performed way better than the systems that were trained on MSA to English corpus.

References:

* Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, Chris Callison-Burch “*Machine Translation of Arabic Dialects”* [*http://www.aclweb.org/anthology/N12-1006*](http://www.aclweb.org/anthology/N12-1006)
* Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh “*Morphological Analysis and Disambiguation for Dialectal Arabic”* [*http://www.aclweb.org/anthology/N13-1044*](http://www.aclweb.org/anthology/N13-1044)