Machine Translation

Course Project Proposal

Weston Feely and Serena Jeblee

February 28, 2012

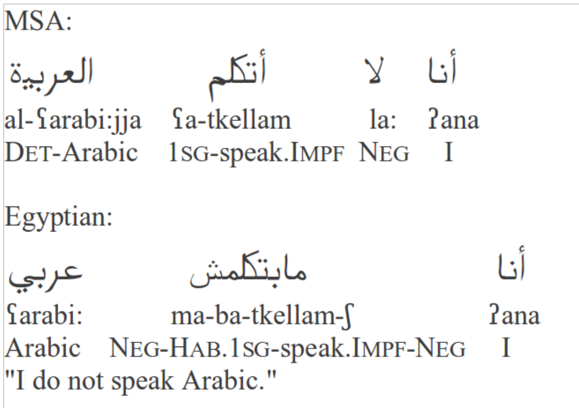
Problem

While there has been much research done on Modern Standard Arabic (MSA) in the machine translation community, there has been little work done on dialectal Arabic. For this class project, we will attempt to address the lack of research in dialectal Arabic MT by building a MT system for Egyptian Arabic, the most spoken dialect of Arabic.

MSA is the standardized variety of Arabic used for formal speech and writing, which is based on millenia of Arabic literary history and is taught in schools. Until recently, MSA was the only written form of Arabic, and so most of the written Arabic corpora are in MSA.

However, with the rise of the internet and social media, people have begun to write in dialectal Arabic, although Arabic dialects have no standardized spelling. In colloquial discourse native Arabic speakers use their local dialect, many of which are low mutual intelligibity with each other and with MSA. Many of these dialects have several million speakers; Egyptian Arabic has about 54 million speakers worldwide, with the majority of these speakers being in Egypt. Additionally, Egyptian Arabic is widely understood in other Arabic-speaking countries because of Egypt's large media influence and entertainment industry.

As a comparison of MSA and Egyptian Arabic, below is a example of the morphosyntactic differences between MSA and Egyptian, for a single sentence meaning "I do not speak Arabic."



MSA:

العربية	أتكلم	لا	أنا
al-ʕarabi:jja	ʕa-tkellam	la:	ʔana
DET-Arabic	1SG-speak.IMPF	NEG	I

Egyptian:

عربي	مابتكلمش	أنا
ʕarabi:	ma-ba-tkellam-ʃ	ʔana
Arabic	NEG-HAB.1SG-speak.IMPF-NEG	I

"I do not speak Arabic."

This is an example of how MSA and Egyptian Arabic differ in their use of a morphological affixes. The same verb for "speak" is accompanied by a negation auxiliary "la:" in MSA, while in Egyptian Arabic, a negation circumfix "ma- -ʃ" performs the same grammatical function. The Egyptian verb prefix "ba-" carries the first person singular agreement features for the verb, but also carries habitual aspect. In contrast, the MSA first person singular verb prefix "ʔa-" has variant spelling and pronunciation, and does not have habitual aspect, since MSA does not have habitual aspect. In order to correctly translate such Egyptian Arabic verbs, these complex morphological affixes must be segemented and translated separately. However, currently there is no freely-available morphological analyzer that can do this for Egyptian Arabic, and popular morphological analyzers for MSA will fail to capture the morphemes like the negation circumfix described above when applied to Egyptian Arabic, which differ drastically from MSA morphemes.

As an example of current state-of-the-art MT for MSA and Egyptian Arabic, below is the Google Translate output of the same sentences as in the above figure.

The Google Translate output for the above MSA sentence is the correct translation:

```
I do not speak Arabic
```

However, the Google Translate output for the above Egyptian Arabic sentence is far worse:

`I Mapetklmh Arabic`

Although, this is due to the fact that the Google Translate system for Arabic is based on MSA and not on dialectal Arabic, like Egyptian Arabic.

Proposed Solution

Our proposed solution for this problem is to use morphological analysis to pre-process our Egyptian Arabic data and then hierarchical phrase-based MT to translate our Egyptian Arabic data into English. We will base our project on a recent study by Zbib et al. (2012), which created a parallel corpus of Egyptian Arabic and Levantine Arabic with English. Zbib et al. selected a training data set of Egyptian Arabic, Levantine Arabic, and MSA in varying proportions, which was then run through MADA (Habash and Rambow (2005)), a MSA morphological analyzer, and then used to train a hierarchical phrase-based MT system. We will re-implement their approach for Egyptian Arabic to English MT, which will serve as our baseline model. We will then make improvements over this baseline by creating our own Egyptian Arabic morphological analyzer, to replace MADA, and we will possibly replace their hierarchical phrase-based MT system with a syntax-based MT system for further improvements.

Implementation Plan

Our baseline MT system will be an implementation of Zbib et al.'s Egyptian Arabic to English MT system, using MADA for morphological analysis and a hierarchical phrase-based model trained on Egyptian Arabic, and possibly MSA Arabic, if this performs better than Egyptian Arabic training data alone. We will test our baseline system on Egyptian Arabic data, and compare our baseline results with Zbib et al.

However, since MADA was made for MSA, we expect MADA will incorrectly segment Egyptian Arabic morphemes for many words. To replace MADA, we will attempt to build a Egyptian Arabic morphological analyzer. We will follow Yang et al. (2007), who made a Iraqi Arabic morphological analyzer using a rule-based approach. We will draw information for our morphological analyzer from Egyptian Arabic grammars. We hope to see some improvement in our overall MT system evaluation, after replacing MADA with a Egyptian Arabic morphological analyzer.

Then, time permitting, we may try to implement a syntax-based SCFG MT system to replace the hierarchical phrase-based MT system of Zbib et al. We hope to see further improvement in our overall MT system evaluation, after implementing this syntax-based model.

Evaluation

Our primary evaluation metric for our MT system will be BLEU, following the use of this evaluation metric in Zbib et al. We will evaluate our Egyptian Arabic morphological analyzer as part of the full MT system, using BLEU. However we may also compare our BLEU scores with METEOR as well, to see if there is any difference in METEOR scores after making our improvements to our baseline MT system.

References

1. Rabih Zbib et al. 2012. Machine Translation for Arabic Dialects. In Proceedings of NAACL 2012. Montreal, Canada.

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

3. Mei Yang, Jing Zheng, and Andreas Kathol. 2007. A Semi-Supervised Learning Approach for Morpheme Segmentation for An Arabic Dialect. In Proceedings of Interspeech 2007. Antwerp, Belgium.