Machine Translation of Egyptian Arabic with Morphological Analysis

Serena Jeblee Weston Feely

Carnegie Mellon

Problem

- Modern Standard Arabic (MSA) is the standardized variety of Arabic used for formal speech and writing. Native Arabic speakers use their local dialects in normal conversation, many of which have low mutual intelligibility with each other and with MSA.

- Egyptian Colloquial Arabic has about 54 million speakers worldwide. Additionally, Egyptian Arabic is widely understood in other Arabic-speaking countries because of Egypt's large media influence and entertainment industry.

- MSA has many NLP resources, while dialectal Arabic does not. Until recently, MSA was the only written form of Arabic, however, with the rise of the internet and social media, people have begun to write in dialectal Arabic.

- MSA and Egyptian Arabic have morphosyntactic differences which are not accounted for in most Arabic MT systems. 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	ʔa-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."

Google Translat

MSA: I do not speak Arabic

Egyptian: I Mapetklmh English

Methods

- Our task is to build an Egyptian Arabic to English MT system. We do this by creating a morphological analyzer to preprocess the data and we use the Moses Decoder to train and test a phrase-based translation system.

- Morphological analysis is done using FOMA [3], a freely available morphological analysis toolkit that uses FSTs to perform the analysis. Because of morphosyntactic ambiguities we may have multiple analyses for a single word, in which case we randomly select an analysis, but we prefer any analysis over no analysis. We compare our analysis to MADA [2], a freely available morphological analyzer designed for MSA which disambiguates different analyses for a single word using an SVM classifier using linguistic features.

- Moses is a freely available MT platform which allows us to build different machine translation systems based on different subsets of the data and levels of analysis.

Data

- Our dataset is bitext corpus of English with Egyptian and Levantine dialectal Arabic which was compiled from web data and translated by workers on Amazon Mechanical Turk [1].

- We use three sets of data: Egyptian only data (EGY), Levantine only data (LEV), and both Egyptian and Levantine data(EGYLEV). Each dataset is randomly separated into training, development and test sets.

- Each group (EGY, LEV, EGYLEV) was trained and tested without morphological analysis, with the MADA analyzer, and with our analyzer built in FOMA.

Results

- BLEU scores from Moses:

	Baseline	MADA	Mish
EGY	13.26	15.28	13.88
LEV	15.80	17.10	16.03
EGY+LEV	16.09	17.51	16.39

Conclusions

- Best results were obtained by using MADA morphological analysis for MT.

- In the future, we plan to improve our FOMA morphological analyzer by training a discriminative classifier using linguistic features, which will better disambiguate multiple morphological analyses for a single word. This will mimic MADA's SVM classifier, which performs a similar function.

- We hope to compare the results of our improved classifier with the results of [1], in future work.

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] Hulden, M. 2009. *foma: a finite-state compiler and library*. In EACL 2009 Proceedings, pages 29-32.