Alla Rozovskaya, Richard Sproat, Elabbas Benmamoun

rozovska@uiuc.edu; rws@uiuc.edu; benmamou@uiuc.edu

Department of Linguistics, University of Illinois, Urbana, USA

Language Modeling of Arabic Dialects

This paper describes several approaches to language modeling of Arabic dialects using Modern Standard Arabic (MSA) data. We build a baseline language model on words and experiment with various techniques of data transformation to account for differences between MSA and Colloquial Arabic. Specifically, we describe three methods of data transformation: morphological simplification (stemming), lexical transductions, and syntactic transformations. We compare the performance of each method with that of the baseline language model. While the best performing model remains the one built using only dialectal data, these techniques allow us to obtain an improvement over the baseline MSA model.

1. Motivation

Processing of Arabic dialects is difficult for several reasons. First, there are not many texts of spoken Arabic available. Second, dialect-specific electronic resources, such as annotated corpora, dictionaries, and parsers have not been developed. Finally, it is hard to develop resources for each dialect, since data transcription is expensive and time-consuming, and there is a whole continuum of Arabic dialects. By contrast, a lot of resources exist for MSA. We therefore wish to determine how one can use MSA data and resources in order to improve language modeling of Arabic dialects. We use the perplexity of test set to evaluate the quality of a language model. Our study thus addresses the following question: is it possible to reduce perplexity of a language model for Colloquial Arabic through use of MSA data?

2. Data

We use the corpus of Egyptian Colloquial Arabic (CallHome), which is a collection of transcribed telephone conversations between native speakers of Egyptian Colloquial Arabic and contains 130K words of training data and 32K words of development data, which we use for testing. We also use the newswire corpus of Modern Standard Arabic (Agence France Presse (AFA) and Al-Hayat (ALH) parts). In addition, we use Al-Hayat part of the Arabic Treebank, which contains data analyzed with Buckwalter Morphological analyzer and annotated with part-of-speech and syntactic information.

3. Experiments

- Baseline language model

The baseline language model built with 130K words of MSA data (AFA) yields a perplexity of 12874.2 . For comparison, a word model trained on the same amount of

Egyptian data gives a perplexity of 184.84. Adding more training data to the MSA model does not help reduce the perplexity.

- Stem language model

The main assumption behind a stem language model is that removing inflections will reduce the amount of morphological discrepancy between the two dialects and will allow us to better model the spoken language with Standard Arabic data. The procedure consists of separating clitics to reduce the number of word types, stripping affixes, and removing short vowels.

We use the LDC Lexicon to extract stems for Egyptian data . The stem models for the AFA data are constructed using Buckwalter Morphological analyzer and SVM package that performs tokenization and POS tagging on Modern Standard Arabic (Diab et al, 2004). Stemming leads to a 50% perplexity reduction (from 12874.2 to 6260.7012) for comparable training data sizes.

- Stem model with lexical transductions

We use a mapping from Egyptian words to MSA equivalents and replace stems in the Egyptian corpus with the stems of their MSA equivalents. This method reduces perplexity further to 2262.31. However, adding more training data does not improve language modeling.

- Word language model with syntactic transformations

We identify frequent syntactic productions in the Al-Hayat part of the Arabic Treebank and apply tree flipping in order to find useful transformations. A transformation is considered useful if its application to the training corpus leads to perplexity reduction. In this manner, we find a certain number of useful transformations, but their effect on perplexity is not significant, especially when compared with the methods described above. For example, the best transformation reduces the perplexity from 12874.2 to 11813.2.

The results suggest that while the techniques help produce a better model, the perplexities are still very large and do not rival the performance of the model trained on the colloquial data alone. Moreover, adding more Standard Arabic data only increases the perplexity. Nevertheless, we believe that these experiments may provide new insights for the problem of modeling Arabic dialects with Standard Arabic. Finally, since we have only experimented with Egyptian Arabic, more research is needed to determine whether the results that the present study has demonstrated hold across other dialects of Arabic.

References:

[1]. Buckwalter Arabic Morphological Analyzer Version 1.0

[2]. Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. *Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks.* Proceedings of HLT-NAACL 2004.

[3]. Katrin Kirchhoff et al. Novel Speech Recognition Models for Arabic Johns-Hopkins University Summer Research Workshop 2002 Final Report

[4]. Mohamed Maamouri, Ann Bies, Hubert Jin, and Tim Buckwalter. 2003. Arabic treebank: Part 1 v 2.0. Distributed by the Linguistic Data Consortium. LDC Catalog No.: LDC2003T06.

[5]. Rambow,Owen, David Chiang, Mona Diab, Nizar Habash, Rebecca Hwa, Khalil Sima'an, Vincent Lacey,Roger Levy,Carol Nichols,Safiullah Shareef. Parsing Arabic Dialects Final Report – Version 1, January 18, 2006 http://www.clsp.jhu.edu/ws2005/groups/arabic/documents/finalreport.pdf