Chuong Ngo - Summary for Alla Rozovskaya’s Colloquium

Generalized Character-Level Spelling Error Correction

**Summary**

Malformed or noisy input data is problematic for natural language processing (NLP) which greatly benefits from clean data. With the increases in the mining of natural language texts generated by users online (suchs as web forum posts, tweets, etc…), the importance for those natural language texts to be clean has increased. An example of the noise that can appear in the input data comes from spelling errors. Spelling errors can lead to increased sparsity and errors when trying to data mine the texts. The Generalized character-Level Spelling Error Correct model (GSEC) is a new model that uses supervised learning to map input characters to output characters with a context, providing a spelling error correction capability. Unlike previous systems that are trained on the words of a language, GSEC does the correction by predicting the next character based upon previous character. Additionally, the one model classifier is used for all common error patterns in the training data consisting of all characters without specific rules. The model also takes into account multiple words for its prediction, allows for the addition of features that may or may not be language-specific, and the model itself is language independent.

In order to achieve the above, GSEC takes an input of *n* characters, including spaces, and transforms that input into a corrected output of m characters one character at a time. For each input character, the character is either passed through without transformation, substituted with another character, deleted, or another character is inserted before the input character. The classifier used is a multi-class SVM and it predicts the action labels, or what transform to apply, for each input character. One restriction on the system is that it only does single-edit actions, ignoring more complex multiple edit actions like multiple insertions or combinations of insertions and substitutions. Training was done using a parallel corpus consisting of aligned erroneous and corrected reference texts for the Egyptian Arabic dialect.

Evaluation on the GSEC model was done by looking at the word-error-rate (WER), the correct-rate (Corr), and the accuracy metric (Acc). For error correction, GSEC was tested with bigram, trigram, and 4-gram N-gram features as well as the maximum likelihood estimate (MLE) model. The accuracy of the various GSEC models achieved accuracies from 90.3% to 92.2%. Using just Ngram features for error correction, GSEC achieved a maximum accuracy of 1.6% over CEC, 2.3% higher correct-rate, and 2.2% lower word-error-rate. The combination of GSEC with a 4-gram feature and MLE for error correction, the improvements in accuracy, correct-rate, and word-error-rate is diminished. When compared to the baseline, the improvements from GSEC + 4-gram + MLE model are greater with accuracy and correct-rate jumping 16.6% and 17.3% respectively while the word-error-rate dropped to about a fourth.

**Reference**

Farra, N., Tomeh, N., Rozovskaya, A., & Habash, N. (2014). Generalized character-level spelling error correction. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL), Baltimore, Maryland, USA*.