Kyle Gorman
Google Inc.

Linguistic insights in text normalization

Abstract: Many speech and language applications, including speech recognition and synthesis, require mappings between "written" and "spoken" (e.g., pronounceable) forms of entities like cardinal and ordinal numbers, dates and times, and the like (e.g., “$328” vs. “three hundred twenty eight dollars”). Collectively, such conversions are known as *text normalization*. Despite substantial progress in applied machine learning, it is still the case that real-world text-to-speech (TTS) synthesis systems largely depend on language-specific hand-written rules. These may require a great deal of development effort and linguistic sophistication and as such, represent substantial barriers for quality control and internationalization.

I first consider the case of number names. I propose two types of computational models for learning this mapping. The first uses end-to-end recurrent neural networks. The second, inspired by prior literature on cross-linguistic variation in number systems, uses an induction strategy based on finite-state transducers. While both models achieve near-perfect performance, the latter model can be trained using several orders of magnitude less data, making it particularly useful for low-resource languages. The latter model is currently being used at Google to develop new number name grammars for hundreds of languages.

I then consider homographs, i.e., words whose pronunciation depends on the intended sense (e.g., *bow*, pronounced as either [boʊ] or [baʊ]). I describe an existing rule-based system and compare it to a novel system which performs homograph disambiguation using simple machine learning. An evaluation of these two systems, using a new, freely-available data set, finds that a hybrid system (using both rules and machine learning) is significantly more accurate than either rules or machine learning alone. This new hybrid system is used on all English TTS traffic at Google.