Leveraging Twitter for Low-Resource Conversational Speech Language Modeling

Aaron Jaech, and Mari Ostendorf

***Abstract*—In applications involving conversational speech, data sparsity is a limiting factor in building a better language model. We propose a simple, language-independent method to quickly harvest large amounts of data from Twitter to supplement a smaller training set that is more closely matched to the domain. The techniques lead to a significant reduction in perplexity on four low-resource languages even though the presence on Twitter of these languages is relatively small. We also find that the Twitter text is more useful for learning word classes than the in-domain text and that use of these word classes leads to further reductions in perplexity. Additionally, we introduce a method of using social and textual information to prioritize the download queue during the Twitter crawling. This maximizes the amount of useful data that can be collected, impacting both perplexity and vocabulary coverage.**

I. Introduction

CONVERSATIONAL speech is very different in style from broadcast news and prepared speeches, so the language models used in automatic speech recognition (ASR) of conversational speech rely on training from speech transcripts, which are more costly to obtain than more formal written text. Thus, data sparsity is typically the limiting factor for language model performance. For many less well-studied languages, there is little transcribed speech available and obtaining additional training data in the target domain is no easy task. An attractive alternative to collecting additional data is to direct that effort towards building models that require less training data. Since little additional work is needed to reuse these models, the benefit of this effort is compounded with each new language thus lowering the barrier for bringing ASR to low-resource languages. While new modeling approaches are valuable, it is the case that even these models benefit from additional data, and researchers have long been exploring mechanisms for using out-of-domain data in combination with a small in-domain corpus to build more robust language models. Different sources of spontaneous speech can help with common words, but these are often not available for less well-studied languages. Even with such data, covering domain-specific vocabulary typically means using written text sources for training language models for ASR. Finding informal text that is useful for modeling conversational speech can be a challenge.

The Internet has been an attractive place to go for researchers looking to expand their training data, as described in the next section. However, if done without care, pulling large amounts of data from the Internet will give no benefit. For example, in work on recognizing English conversational speech in broadcast talk shows, the Google n-grams provided no benefit to perplexity or word error rate [1]. Here, we instead

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0014. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

A. Jaech and M. Ostendorf are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA e-mail: {ajaech,ostendor}@uw.edu.