Title: Corpus Creation for Disfluency Research

Author: Stephanie M. Strassel
Affiliation: Linguistic Data Consortium, University of Pennsylvania

Abstract

Large-scale linguistic corpora are increasingly available in an ever-growing number of languages, providing a rich pool spontaneous speech data for the investigation of disfluencies and related discourse phenomena. Much of the empirical data required for disfluency research is collected, annotated and analyzed by individuals or individual research groups. While pilot investigations by individual researchers are essential to achieving a basic understanding of the structure of disfluencies, creating vast amounts of training data in order to train models for automatic speech processing requires resources well beyond the reach of a single research site.

This paper reports on work in progress under the DARPA EARS Program to create large-scale annotated corpora to support disfluency research, including fillers (filled pauses, discourse markers, asides/parentheticals and editing terms) and edits (repetitions, revisions and restarts). The fundamental goal of EARS is a significant reduction in word error rate (WER) for automatic speech recognition systems. By detecting and characterizing disfluencies and other "metadata" phenomena in the speech stream, systems should achieve an overall improvement in WER.

Data annotated for disfluencies and related phenomena under the EARS Program includes hundreds of hours of English conversational telephone speech, plus tens of hours of pilot data in Mandarin Chinese and Egyptian Arabic. Though created within the context of EARS, the data is slated for eventual publication, making it available to the wider research community. In addition to describing the EARS metadata corpora, the paper covers the process of corpus creation. Details of the annotation process are presented, including the challenge of creating annotation guidelines that allow for multiple, non-expert annotators to achieve high levels of inter-annotator consistency while maintaining maximal efficiency. Issues of annotator training, annotation tools and quality assurance measures are described, and the paper concludes with a discussion of some of the difficulties of adapting the English-based guidelines and processes to new languages.