Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 23 – 29

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Mismatched Crowdsourcing based Language Perception for Under-resourced Languages

Wenda Chen[a,b,*], Mark Hasegawa-Johnson[a], Nancy F. Chen[b]

[a]*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 306 N. Wright St. Urbana, IL 61801-2918, Illinois, USA*

[b]*Human Language Technology Department, Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632*

Abstract

Mismatched crowdsourcing is a technique for acquiring automatic speech recognizer training data in under-resourced languages by decoding the transcriptions of workers who don't know the target language using a noisy-channel model of cross-language speech perception. All previous mismatched crowdsourcing studies have used English transcribers; this study is the first to recruit transcribers with a different native language, in this case, Mandarin Chinese. Using these data we are able to compute statistical models of cross-language perception of the tones and phonemes from transcribers based on phone distinctive features and tone features. By analyzing the phonetic and tonal variation mappings and coverages compared with the dictionary of the target language, we evaluate the different native languages' effect on the transcribers' performances.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Mismatched Crowdsourcing, Speech Recognition, Low Resource Language, Speech Perception

1. Introduction

In many languages, it is hard to find training data for automatic speech recognition, because it is hard to hire native transcribers. Mismatched crowdsourcing[1,2,3] bypasses the need for native transcribers by recruiting

* Corresponding author. Tel: +1-217-751-2069.
E-mail address: wchen113@illinois.edu

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.025

transcribers who don't speak the language. Transcribers write what they hear as if it were nonsense speech in their own language; their transcriptions are then decoded using a noisy-channel model of second-language speech perception. All previous published studies of mismatched crowdsourcing used English-speaking transcribers, recruited on Amazon Mechanical Turk. By decoding their transcriptions using weighted finite state transducers, it has been demonstrated that ASR can be trained with reduced phone error rate compared with the multilingual and semi-supervised approaches for low-resourced languages[4].

Mismatched crowdsourcing, as proposed in[2], requires a model of the misperception of phonemes because of mismatch between the speech language and the transcriber language. Previous studies have examined several different speech languages[4], but have always assumed English to be the transcriber language. This paper generalizes mismatched crowdsourcing by requesting transcriptions from native speakers of a tonal language, specifically, Mandarin Chinese. Mandarin speakers create transcriptions of both the phonetic and tonal content of utterances, using Pinyin orthography. With the introduction of the transcriber's language as an additional variable, the Bayesian model used to decode mismatched transcription can be extended. Prior work has developed techniques to merge the manual transcripts T into a probabilistic distribution over cleaned representative transcripts[3,4]. We focus on improving this distribution model by adding the new dependent factor of transcribers' native languages.

2. Mismatched Crowdsourcing Experiments

For our experiments, we choose Vietnamese and Cantonese to be our under-resourced languages and we employ two sets of crowd workers with different language backgrounds: 10 random English speakers for each sentence (employed on Amazon Mechanical Turk) and 6 consistent Mandarin speakers for all sentences (mainly employed on Upwork (www.upwork.com)). Each crowd worker listens to a short speech clip in Vietnamese or Cantonese and provides a transcription that is acoustically closest to what they think they heard. The transcriptions from the English speakers are in English (mostly in the form of nonsense syllables and not corresponding to valid English words) and the Mandarin speakers use the Pinyin alphabet. Vietnamese and Cantonese speech samples were downloaded from the Australian Special Broadcasting Service (http://www.sbs.com.au/podcasts/yourlanguage). Bumpers and non-speech audio were discarded; the remaining speech was cut into overlapping one-second segments, without regard for word boundaries. Speech totalling 1 hour was transcribed in both Vietnamese and Cantonese. Native speakers of both Vietnamese and Cantonese were recruited to provide reference transcriptions in each of these two languages, so that the results of mismatched crowdsourcing can be evaluated with respect to native transcriptions[5]. With the considerations and normalization of the intra and inter worker agreement, we will look at phone variations and tone variations in the transcriptions.

Transcribers recruited for this study differ in two important ways. First, Mandarin transcribers may be able to transcribe Cantonese and Vietnamese more accurately than English transcribers, because the syllable structures of Cantonese and Vietnamese resemble that of Mandarin more than that of English. Second, however, there are two different crowdsourcing markets. English-speaking transcribers were recruited anonymously on Mechanical Turk, so our ability to apply quality control was quite limited. Most Mandarin-speaking transcribers (all but for 50 sentences, described in the next paragraph) were recruited on Upwork. Each of the six Mandarin transcribers was given the full hour of speech, in both Vietnamese and Cantonese, thus they are able to check for consistency of consecutive sections. Furthermore, Upwork transcribers were hired by name and their working history with profiles, therefore they have incentive to provide high-quality transcriptions so that they will be hired in the future.

Next we will describe a comparison of the Mandarin (Pinyin and tone) and English (nonsense words) mismatched crowdsourcing data collected from Upwork and Mechanical Turk. For this comparison purpose, we collected 50 additional sentences of testing data from Mechanical Turk transcribing Vietnamese and Cantonese in Mandarin. From the Mandarin data collected in both sources, the top three most common phone confusions between Upwork and MTurk data are listed in Table 1. They are largely dependent on the transcribers for both cases while te common substitutions in MTurk data are more random and unpredictable according to the literature of language perceptions. In general, Mandarin Pinyin system helps limiting the free choice of the phonetic symbols. The Mandarin data from MTurk has more uncontrolled usage of the Pinyin symbols and varies much more in the word coverage. The English data from MTurk suffers from much more variations of the symbol usage, and even includes many cases of syllable insertion and deletion, the normalization and filtering of which require significant post-processing[3].