Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Transcribing and Annotating Speech Corpora for Speech Recognition: A Three-Step Crowdsourcing Approach with Quality Control

Annika Hämäläinen[1,2], Fernando Pinto Moreira[1], Jairo Avelar[1], Daniela Braga[1,2], Miguel Sales Dias[1,2]
[1]Microsoft Language Development Center, Lisbon, Portugal
[2]ADETTI – ISCTE, IUL, Lisbon, Portugal
{t-anhama, t-fpinto, t-jairol, dbraga, Miguel.Dias}@microsoft.com

Abstract

Large speech corpora with word-level transcriptions annotated for noises and disfluent speech are necessary for training automatic speech recognisers. Crowdsourcing is a lower-cost, faster-turnaround, highly scalable alternative for expert transcription and annotation. In this paper, we showcase our three-step crowdsourcing approach motivated by the importance of accurate transcriptions and annotations.

Automatic Speech Recognition (ASR) is technology that translates a speech signal into a sequence of words. Well-known ASR-driven applications include, for instance, voice search (e.g. Bing voice search) and the intelligent personal assistants of smartphones (e.g. Apple's Siri). To translate a speech signal into a sequence of words, speech recognisers use acoustic models (AMs), statistical abstractions that model the discretised acoustic properties of the observed speech sounds, or phonemes, of a given language. However, before a recognition task can be performed, the AMs must be trained using a large corpus of language-specific speech.

To be able to train AMs that model the acoustic properties of phonemes as accurately as possible, the corpus used for training the AMs must contain orthographic (word-level) transcriptions of the spoken material. A pronunciation dictionary is then used to identify the underlying phonemes whose acoustic properties in the speech signal are used for training the AMs. However, speech corpora do not usually only contain carefully spoken speech; the recorded audio files often also contain hesitation phenomena, such as phoneme lengthening (e.g. *theeeee book*), filled pauses (e.g. *ah*, *um*) and false starts (e.g. *bo- book*), speaker noises (e.g. lip smacks), non-human noises (e.g. music), and damaged (e.g. mispronounced) words. To prevent such audio events from contaminating the statistical phoneme models and negatively affecting speech recognition performance, they must be annotated and trained separate AMs for.

Speech corpora have traditionally been transcribed and annotated by teams of expert transcribers that have specifically been trained for the task. However, due to the slowness and high cost associated with expert transcription and annotation, there is considerable interest in using crowdsourcing (CS) for the work. Past efforts have shown that the quality of non-expert transcription can approach that of expert transcription, while substantial savings can be achieved both in terms of time and money (Parent and Eskenazi 2011). However, to produce high-quality data, two aspects seem particularly important: 1) breaking the work down into cognitively undemanding subtasks and 2) quality control (Parent and Eskenazi 2011). In this paper, we present our approach to dealing with these two issues: breaking the work down into three simple steps, or human intelligence tasks (HITs), and taking various measures to improve the accuracy of the produced transcriptions and annotations.

Three-Step Approach

We are using Microsoft's Universal Human Relevance System (UHRS) CS platform for our work. UHRS is a marketplace that connects HITs with a large pool of workers, or judges, who are hired by third-party vendors.

To showcase our approach, we use a corpus of read speech. In practice, this means that the initial transcriptions are the prompts that the speakers were asked to read out. However, before using the prompts in UHRS, we process them automatically to eliminate punctuation, to convert unnormalised abbreviations and number expressions (e.g. dates) to their written forms, and to lowercase words other than proper nouns, acronyms etc. Our approach can also be used with spontaneous speech; in this case, the initial tran-

scriptions can be generated using automatic speech recognition.

Step 1: Identifying Problematic Utterances

The transcription work starts by identifying transcriptions that require orthographic changes, and audio files that do not contain any intelligible speech. To this end, in Step 1, we ask the judges to listen to audio files (with each of them expected to contain one utterance corresponding to one speaker prompt) by using an audio player embedded in a HIT App, to read the corresponding transcriptions, and to answer a Yes/No question. When selecting "No", the judges must motivate their answer by ticking one or both of two check boxes: 1) the transcription is somehow different from the utterance and 2) the audio quality is not acceptable (e.g. there is so much background noise that you cannot hear the speech properly).

Step 2: Correcting Transcriptions

In Step 2, we process all the transcriptions that the judges marked as requiring orthographic changes in Step 1. In practice, we ask the judges to listen to the audio files, to read the corresponding transcriptions that are now presented in a text box, and to add, delete and substitute words in the text box such that the resulting transcriptions match the utterances in the audio files. We encourage the judges to switch on automatic spell checking so that it will be easier for them to spot any spelling errors that they might make. After Step 2, we automatically look for and correct common spelling errors in the transcriptions.

Step 3: Annotating Transcriptions

In Step 3, we again process all the transcriptions, applying the orthographic changes made in Step 2. We now ask the judges to listen to the audio files, to read the corresponding transcriptions, which are presented in a text box, and to annotate all clearly audible hesitation phenomena, speaker noises, non-human noises, and damaged words. They can do this by moving the cursor to the right place in the text box and by clicking one of the buttons that correspond to the different types of annotations, or by deleting a word and replacing it with an annotation.

Quality Control

We are taking several measures to ensure high-quality transcriptions and annotations. First, our vendor only hires judges that are native speakers of the language in question. Second, before judges are allowed to start working on a HIT App (or Step), they are presented with the transcription/annotation guidelines for that Step and must pass a prequalification test that is identical to the HIT App but provides them with feedback on their correct and incorrect answers and, hence, trains them for the work. Judges that fail the prequalification test three times are not allowed to work on that particular HIT App.

In the transcription/annotation guidelines, which the judges can access at any given time during the prequalification test and the actual work, we highlight the importance of accuracy. In practice, we monitor the judges' performance by regularly inserting *gold standard* utterances (utterances whose transcriptions and/or annotations we know) into our HIT Apps. Similarly to the prequalification tests, the judges are provided with feedback on their answers and are, therefore, also being trained during the work.

When the judges are not sure how to transcribe/annotate an utterance, we ask them to move on to the next HIT without providing an answer. That way, we can reallocate such HITs to other judges until they are answered.

Finally, we improve the quality of our data by asking several judges to complete the same HITs. In Step 1, we use a consensus approach. Up to three judges complete the same HIT; we stop submitting the HIT to the HIT App as soon as two judges have given the same answer. In Steps 2 and 3, we ask one judge to correct or annotate a transcription and then submit the new transcription to another judge for possible corrections. This is different from previous work, in which the same utterance is transcribed by several judges and the transcriptions are then merged (Parent and Eskenazi 2011), but similar to what is sometimes done when using expert transcribers (e.g. Cucchiarini et al. 2008). If it turns out that two passes are sufficient for getting high-quality transcriptions, our approach will be cheaper than merging.

Future Work

We are currently employing the developed HIT Apps to transcribe and annotate a corpus of read European Portuguese elderly speech (Hämäläinen et al. 2012). The quality of the resulting non-expert transcriptions and annotations will be compared with that of expert transcriptions and annotations; the results of this work will be presented elsewhere.

References

Parent, G., and Eskenazi, M. 2011. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In: Proc. *Interspeech*. Florence, Italy.

Cucchiarini, C., Driesen, J., Van hamme, H., and Sanders, E. 2008. Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: The JASMIN-CGN Corpus. In: Proc. *LREC*. Marrakech, Morocco.

Hämäläinen, A., Pinto, F., Dias, M., Júdice, A., Freitas, J., Pires, C., Teixeira, V., Calado, A., and Braga, D. 2012. The First European Portuguese Elderly Speech Corpus. In: Proc: *IberSPEECH*. Madrid, Spain.