End-to-End Crowdsourcing Approach for Unbiased High-Quality Transcription of Speech

Michael Levit, Shuangyu Chang, Omar Alonso, Anil Kumar Yadavalli
Microsoft Corporation

Abstract

We present an end-to-end implementation of a crowdsourcing speech transcription pipeline that aims at achieving multiple goals including high transcription fidelity, minimal bias towards machine-generated recognition hypotheses and low cost. Our approach consists of two stages: unassisted transcription and variant selection. Each stage is realized as an iterative process where opinions are solicited from judges as long as no reliable decision regarding final utterance transcription can be made. Acknowledging possible ambiguity of the hypothesis space, our final consensus hypotheses can comprise several alternative transcriptions for each utterance merging them into a single word confusion network. Using lexicographic transcription task for Microsoft Cortana, we show that our approach produces low cost transcriptions that are superior even to the professional transcriptions in terms of exposure bias, accuracy and latency.

Being able to generate reliable speech transcriptions is crucial for achieving success in many tasks of human machine communication. The importance of correct references is evident for training stochastic models (e.g. Language Models) that suffer from presence of noisy samples, but also for testing, where even small reference bias can unfairly penalize a superior model. While commonly assumed unambiguous and incontestable, such references are often anything but that. Apart from the known fact that transcribers, and in particular unskilled crowd judges, can have insufficient expertise (e.g. be unfamiliar with certain named entities) or make avoidable mistakes (e.g. typos), the field often grapples with true ambiguity (*"call chris"* or *"call kris"*) by virtue of lacking discourse context and/or not being able to get in the mind of the utterance originator. In some cases, speakers could even have difficulties transcribing their own speech (e.g. was it *"you are right"* or *"you're right"*). This has several consequences. First, as a single transcription might be unattainable for each utterance, the resulting reference can incorporate a number of (possibly weighted) alternatives. Second, transcribers should be assisted by some knowledgeable automated system (e.g. production Automatic Recognition System, ASR, that is aware of millions of rare words and names). However, used on its own, this extension is risky, as it primes judges with an often plausible hypothesis, making her give the ASR benefit of the doubt and trust it more than it deserves, especially when acoustic conditions are adverse or homophonic alternatives are possible. This backfires when a new ASR is developed and evaluated for deployment in place of the production one. Our experiments show that in realistic cases, up to 10% of relative Word Error Rate (WER) improvement due to the new candidate ASR can be masked by the transcription bias of transcribers primed by exposure to the production system recognition results.

In the following, we present a novel crowdsourcing transcription pipeline recently introduced at Microsoft for high volume lexical transcription of speech that addresses all of the aforementioned points. We explain individual steps of the transcription pipeline listing optional refinement techniques and illustrate the improvements that this pipeline achieves with respect to the baseline approach of assisted professional transcriptions.

System Description

The high level representation of our system is shown in Figure 1. The pipeline starts with providing each of the utterances with two or more alternative automatic recognition results (1). Ideally, the systems should be diverse but of comparable quality. In practice, while deciding whether to deploy a new ASR instead of the production ASR, recognition hypotheses from both systems can be used. During the first transcription stage, only audio is played to the judges, one judge at a time, and lexical transcriptions are solicited (2). As new suggestions arrive for an utterance, we compile them into a single cumulative distribution (3) of alternative transcription hypotheses and their scores derived from recognition confidences (for ASR) and judge reliability estimates. The value of its normalized entropy is then used to decide whether a single transcription should be promoted to be the sole output (4a), or the currently harvested hypotheses should be passed to the next stage for selection (4b), or more iterations (judges) are required for this utterance (4c).

If the selection stage is required for an utterance, we take the unique hypotheses obtained for it (from human judges or ASR) and present them to a different pool of judges in random order without disclosing their sources. Given utterance audio, and the above list of alternatives, the selection stage

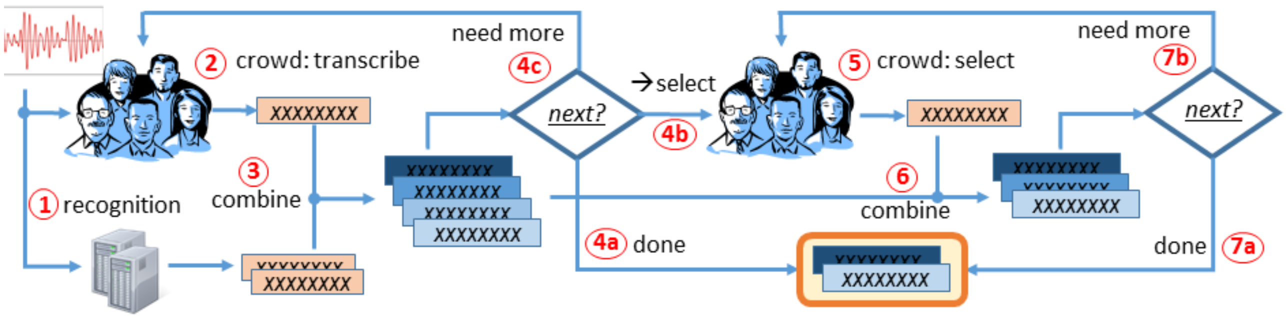

Figure 1: *Iterative two-stage crowd-transcription pipeline.*

judges are asked, one at a time, to pick the best alternative or provide a new one if none is deemed correct (5). At each iteration, the selected transcription hypothesis is folded into a distribution of weighted alternatives (6), and its normalized entropy is evaluated causing the system to either stop iterations and present one or several selected hypotheses as the final answer (7a) or continue asking more judges (7b). In the above, we require that no hypothesis is returned unless at least one human judge wrote it or selected it. Instrumental in producing distributions of unique transcription hypotheses are ratings associated with the individual judges that are based on several factors such as their agreement with peers. The task difficulty is illustrated by averaged Cohen $\kappa \approx 0.2$ estimated for utterances that were forced into the selection stage. Overall, the two-stage approach is reminiscent of the Fix-and-Verify strategy employed in (Bernstein, M. S. et al. 2010).

The entire pipeline is implemented within Microsoft stack and centered around Microsoft crowdsourcing platform UHRS (Patel 2012). The advantage of our approach consists in introducing ASR's vast lexical knowledge into the annotation pipeline while minimizing transcriber bias, especially when several ASR systems contribute their recognition results. In addition, a number of special techniques turned out to be beneficial from accuracy and cost perspectives. First, we allow judges to mark utterances as not containing any device-directed speech in targeted language. We also allow them to report very difficult cases that would then be dismissed if sufficient evidence as to their difficulty is collected. A number of short-cut rules bypassing entropy estimation was established (e.g. accept a single hypothesis in the first stage if the first three human transcribers suggested the same transcription). As it is common in the field, our pipeline requires judge candidates to pass qualification test, and the judges are then periodically educated and tested with special samples for which a single correct transcription is known. Finally, we incorporate real-time Bing spell checker into the pipeline that processes each hypothesis on-the-fly and offers the judge to have a second look when alternatives are returned.

Results

There are several objectives that our system is designed to accomplish. One of them is to combine high quality accuracy with unbiased judgments. To evaluate, two experiments have been conducted. A set of 2000 randomly selected utterances from Cortana domain was selected and transcribed using our new pipeline (only the highest scoring transcription was preserved) as well as via ASR-assisted single-professional-opinion (legacy) pipeline. Then, for the 232 utterances where the two results disagreed, an independent group of professional transcribers was asked to select the best alternative given the audio. In the end, 50% of cases were resolved in favor of the crowdsourcing approach, 40% in favor of the legacy pipeline, and the rest was deemed of equal quality. In a separate experiment, we computed WER difference between two competing ASR systems: once w.r.t. legacy and another time w.r.t. crowdsourcing transcriptions. This difference was correlated with results of a direct side-by-side comparison where independent judges were asked to choose one ASR result over another given audio. Better correlation indicates smaller transcription bias and indeed, on the 496 examples where the two ASRs differed, the crowdsourcing approach produced Pearson coefficient of 0.76, whereas legacy pipeline only 0.68. Finally, we looked at the actual WER numbers according to the two transcription methods. While they exhibit impressive agreement on the ASR that was not used to assist transcribers in the legacy pipeline (11.38% vs 11.34%), on the ASR that was used in the legacy pipeline, the crowdsourcing WER was 9.4% and legacy pipeline WER 8.4%, which amounts to more than 10% relative, quantifies the effect of priming bias and illustrates how it is eliminated in our pipeline.

Conclusion

We have presented a production grade crowdsourcing pipeline that has proved to produce high-quality unbiased and fast transcription of spoken utterances. The process is inexpensive as it currently requires only 3 judgments per utterance on average, and it has proven to outperform professional single-opinion transcriptions on a number of metrics.

References

Bernstein, M. S. et al. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proc. of the 23 Annual ACM Symposium on User Interface Software and Technology*.

Patel, R. 2012. http://research.microsoft.com/en-us/um/redmond/events/fs2012/presentations/Rajesh_Patel.pdf.