ACCURATE TRANSCRIPTION OF BROADCAST NEWS SPEECH USING MULTIPLE NOISY TRANSCRIBERS AND UNSUPERVISED RELIABILITY METRICS

Kartik Audhkhasi, Panayiotis Georgiou, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA, U.S.A.
Email: audhkhas@usc.edu, {georgiou, shri}@sipi.usc.edu

ABSTRACT

Professional manual transcription of speech is an expensive and time consuming process. This paper focuses on the problem of combining noisy transcriptions from multiple non-expert transcribers, where the quality of work from each worker varies. Computing transcriber reliability is a difficult task in the absence of gold standard reference transcripts. Three simple metrics for quantifying this reliability without using a gold standard are proposed. We create a database of 1000 Mexican Spanish broadcast news audio clips transcribed by five transcribers each through Amazon Mechanical Turk. Combination of multiple noisy transcripts using these reliability scores improves the word error rate of the combined transcript with respect to the LDC gold standard by 8 % relative, and the sentence error rate by 4.1 % relative, when compared with a combination without any reliability information.

Index Terms— Speech transcription, evaluator reliability, crowd sourcing

1. INTRODUCTION

In every supervised learning pattern classification problem, each training example must be associated with a class label. Many spoken language processing systems, like speech recognizers, fall into this category, and require the availability of reference ranscriptions. Clean and accurate transriptions are critical to both training and testing system performance. Being such a crucial part of the training data, the transcriptions are typically prepared by trained evaluators, who are not only acquainted with possible variabilities in the audio (for example, accent, speaking rate, background noise), but also are knowledgeable about the pronunciation and orthography of different words and of transcription conventions of the specific dataset. All these requirements make transcription an extremely time-consuming and expensive process.

In recent years, there have been many efforts in both the machine learning and speech/language processing community towards doing away with the requirement of having expert annotators for labeling the data. All these works utilize multiple non-expert annotators to perform the labeling. A combination of these multiple labels is expected to be closer to an expert labeling than the individual labels. Amazon's Mechanical Turk (MTurk) is a web-service which allows *workers* from all over the world to perform some *Human Intelligence Tasks* (HITs) required by *requesters*. The requester creates a simple web-interface through which workers can submit their answers. If the requester deems the quality of the answers satisfactory, he pays the workers a pre-specified amount of money per HIT.

Many recent works in natural language processing have focussed on utilizing this crowd-sourcing web service for various tasks. Snow et al. [1] use MTurk for getting non-expert annotations for five tasks - affect recognition, word similarity, recognizing textual entailment, temporal event recognition and word sense disambiguation. They show high correlation between the gold standard and non-expert labels for all tasks. They also demonstrate that for affect recognition, using labels from non-experts for training the system gives similar performance as when using the gold standard labels. Callison-Burch [2] uses MTurk for manually evaluating machine translation quality, and finds that the evaluations not only have high agreement with the gold-standard evaluations, but also correlate more strongly than the BLEU score. Lambert et al. [3] create a linguistic plausibility dataset by having randomly generated sentences from a N-gram Language Model (LM) annotated by workers in MTurk. A workshop in NAACL HLT-2010 asked researchers to submit papers regarding annotation of speech and language data using MTurk, subject to a maximum budget of $100. Accepted papers covered a wide variety of topics, ranging from paraphrasing for English-Arabic machine translation [4] to building parallel corpora for machine translation systems [5].

Previous work has also focussed on using MTurk for speech transcription. Marge et al. [6] present a study about transcription of audio consisting of route instructions for robots using MTurk. They also show that combination of multiple non-expert transcriptions using the ROVER algorithm [7] significantly reduces the WER of combined transcripts with respect to the gold standard. A recent work by Roy et al. [8] shows that the force alignment score from an automatic speech recognition (ASR) system is reasonably correlated with the inter-annotator agreement. They also hypothesize that one can predict the transcription difficulty of an audio clip by using ASR scores. This is demonstrated by a significant agreement between the ASR score and a measure reflecting the time consumed by the annotator while transcribing the audio.

This paper also focusses on the problem of transcribing speech by multiple non-expert transcribers. The novelty of our work is two-fold. First, we propose three simple metrics which capture the reliability of a transcript. All these metrics can be computed without the knowledge of the gold standard transcript, and hence are unsupervised in this sense. Second, we show that incorporating this reliability information while combining multiple non-expert transcriptions gives significant improvement in the quality of the combined transcription, as measured by the WER with respect to a gold standard.

This paper is organized as follows. Section 2 describes the

This work was supported by the NSF, DARPA and Army.

database collected through MTurk. Section 3 introduces the three reliability metrics. Section 4 discusses experiments conducted with unweighted and reliability-weighted combination of multiple noisy transcripts using ROVER. We end the paper in section 5 with some conclusions and directions for future work.

2. DESCRIPTION OF THE DATABASE

The database consists of 1000 audio clips from the 1997 Spanish Broadcast News Corpora (HUB4-NE) [9], available through the Linguistic Data Consortium (LDC). These clips contain recordings of broadcast news from Televisa, Univision and Voice Of America (VOA). The gold standard reference transcripts of this database are also available through LDC. We should note that despite the extensive LDC rules of transcription and quality checks, these reference transcripts can contain errors, and can be of different convention to an untrained MTurk transcriber. However, the untrained worker is not necessarily wrong and may simply follow a different transcription scheme. For example, "décimo novena" and "décimonovena" both refer to "nineteenth" in Spanish, but MTurk workers overwhelmingly prefer the former, and LDC, the latter. Finally, being recordings from broadcast news, the audio posseses appreciable background noise, including music and vehicular noise.

For transcription of the audio clips in MTurk, the workers had 2 minutes per HIT, and were given the following key instructions:

"*Transcribe every word you hear in the audio clip using correct Spanish characters and words. Write all numbers in words, as they are spoken, not in digits. If a word is unintelligible, write your best guess of the word in double parantheses - "(())". In case of fragment or partial words, add a "-" after the word. In case of acronyms pronounced as a single word, place a "~" before the word. For acronyms pronounced letter-by-letter, omit the "~".*"

In addition to transcribing the speech, the workers were asked to answer three questions regarding the transcription: background music/noise (yes/no), difficulty of transcribing (1 to 3, with 1 being very easy and 3 being extremely difficult) and confidence of transcription (1 to 3, with 1 being not confident at all, and 3 being totally confident). We didn't use the answers to these questions in this paper, and will address them in a future work.

5 HITs were released for each audio clip, and the reward of each HIT was set at $0.04. This reward was chosen by taking a cue from the experiments conducted in [6]. A small pilot experiment suggested that HITs open to workers only from U.S.A were completed at a much slower rate than those open to the rest of the countries. Hence, it was decided to open the 5000 HITs to all countries except U.S.A. Spurious/fraudulent transcriptions were screened by manual examination, and coherence between the transcripts obtained from different workers. A total of 19 HITs out of 5000 were rejected, and put up again for evaluation. Most of the rejected HITs contained translation of the audio in English, blank transcriptions or digits for representing numbers. An attempt was made to keep the cleaning to a minimum, since such manual post-processing becomes very tough on large databases.

A total of 19 workers annotated the entire database, and 5000 HITs were completed in 5 days. Figure 1 shows the a plot of the number of HITs completed by each worker in decreasing order. Figure 2 shows a histogram of the time taken (in seconds) per HIT. The mean completion time per HIT was 34.5 sec.

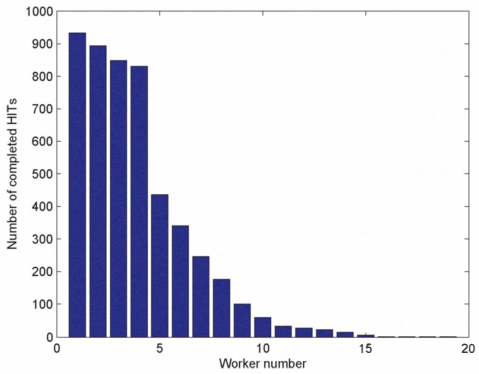

Fig. 1. Number of HITs completed per worker in descending order. Each worker can complete a maximum of 1000 HITs.

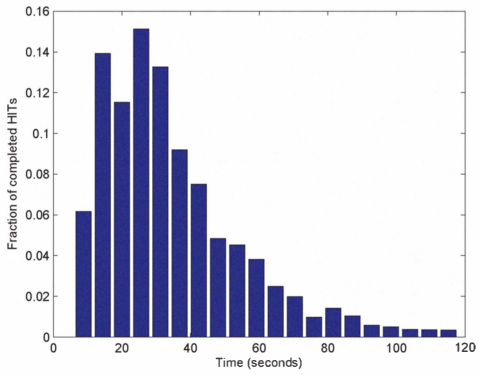

Fig. 2. Histogram of the time taken (in seconds) per completed HIT. A maximum of 2 minutes were allotted for completing each HIT.

3. UNSUPERVISED TRANSCRIPTION QUALITY RELIABILITY METRICS

The ROVER algorithm for combination of multiple ASR outputs starts by performing a dynamic programming-based alignment of the transcripts, followed by a vote in each confusion bin [7]. In its simplest form, the voting process chooses the most frequently occurring word hypothesis in each bin. However, it has been shown in [7] that taking a convex combination of the word frequency and a word confidence score during the scoring process gives an even greater improvement in the WER of the combined transcript with respect to a gold standard transcript. Generating ASR hypotheses with confidence scores is a well-studied problem, and most ASR confidence scores are based on computing the posterior probability of word hypotheses. However, in our case, we are interested in the mismatch between the non-expert transcription and the audio/gold-standard transcription. Since the gold standard transcription is not available, we hypothesize that the word frequency-based ROVER transcription can be used as a proxy for the unavailable gold standard transcription in estimating worker and transcript reliability scores. This idea forms the basis for our three reliability features.

Let S be the number of audio clips in the database (S = 1000 in our case). Let W_s denote the set of workers which have transcribed a sentence $s \in \{1, ..., S\}$. Let t_{sw} be the transcript for a sentence s by

a worker $w \in W_s$. We first combine the transcripts using ROVER without any reliability information, i.e. just on the basis of word frequency in each confusion bin. Let the output of such a combination of $\{t_{s1}, ..., t_{s|W_s|}\}$ be denoted by r_s^0, and we use it as a proxy for the gold-standard transcription for s. Next, we compute the following reliability metrics for t_{sw}:

1. Normalized force-alignment score: We assume the presence of acoustic models for Mexican Spanish, trained offline on some standard corpus. Given these acoustic models, for each s, we compute the total force alignment log-likelihoods of the ROVER reference, r_s^0, and the individual non-expert transcriptions, t_{sw}. Let these be denoted by $L(r_s^0)$ and $L(t_{sw})$ respectively. We normalize these scores by the total number of words in r_s^0 and t_{sw} to get $\bar{L}(r_s^0)$ and $\bar{L}(t_{sw})$. The normalized force-alignment reliability score of t_{sw} is defined as:

$$a_1(s,w) = \frac{\bar{L}(r_s^0)}{\bar{L}(t_{sw})} \quad (1)$$

   Since r_s^0 is expected to be closer to the unobserved gold standard transcription than t_{sw}, its force alignment score is expected to be greater (less negative), and hence $a_1(s,w)$ will lie between 0 and 1. However, there is no theoretical guarantee that this will always happen.

2. Local WER-based reliability: Since we assume r_s^0 to be a proxy for the gold-standard transcripts, our second reliability feature is based on the WER between r_s^0 and t_{sw}:

$$a_2(s,w) = 1 - WER(r_s^0, t_{sw}) \quad (2)$$

3. Global WER-based reliability: We note that the local per-utterance WER may vary appreciably from one utterance to another, and may not reflect the overall reliability of the worker. Thus, we compute a global worker-specific WER using $\{r_s^0\}_{s\in\{1,...,S\}}$ as reference. The worker reliability is thus:

$$a_3(w) = 1 - \frac{\sum_{s:w\in W_s} WER(r_s^0, t_{sw})}{|\{s : w \in W_s\}|} \quad (3)$$

   where $|\{s : w \in W_s\}|$ denotes the number of sentences which worker w transcribed.

Our overall reliability score for t_{sw} is computed by taking a convex combination of the above three scores.

$$a(s,w) = \beta_1 a_1(s,w) + \beta_2 a_2(s,w) + (1-\beta_1-\beta_2)a_3(w) \quad (4)$$

where $\beta_1, \beta_2 \geq 0$ and $\beta_1 + \beta_2 \leq 1$. These parameters are tuned on a development set using an exhaustive grid search.

4. EXPERIMENTS AND RESULTS

For computing the force alignment-based reliability metric, we trained continuous observation, triphone acoustic models in Sphinx-3 [10] using 26 hours of Mexican Spanish audio from the HUB4-NE broadcast news corpus. This audio was separate from the 1000 clips put up on MTurk for transcription. The WER of the models when tested on a held-out set from the same database was 32%, using a trigram LM trained in SRILM [11] on the LDC transcripts corresponding to the 26 hours of Mexican Spanish data.

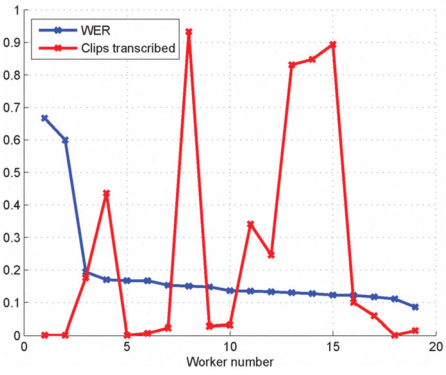

Fig. 3. $\frac{WER}{100}$ and number of clips transcribed by workers as a fraction of 1000, in decreasing order of WER, with respect to gold-standard transcripts

4.1. Baseline ROVER performance

For the baseline, we combined all five non-expert transcriptions of each audio clip using the `avgconf` method in SCTK's [12] implementation of ROVER. This voting method uses a convex combination of word frequency and confidence scores for computing the overall score. We set the tradeoff parameter $\alpha = 1$, thus giving 0 weight to the confidence scores. Figure 3 shows the WER (divided by 100) for all 19 workers in decreasing order, computed with respect to the gold-standard LDC transcripts, and the number of clips transcribed as a fraction of 1000. As we can observe, most of the workers have a WER ranging between 10 to 20 % (0.1 to 0.2 on the vertical axis). Two workers that transcribed just 1 clip have WER exceeding 60 %, but it was ensured by manual inspection that those transcriptions were not spurious.

The set of 1000 audio clips was randomly split into a testing and development set of 800 and 200 clips respectively. No worker partition was done between the testing and development sets. For evaluation of the standard (unweighted) ROVER algorithm, we combined the five worker transcripts for each sentence without any reliability information, i.e. using $\alpha = 1$ in the `avgconf` method in ROVER. Upon comparing with the LDC reference transcripts, the WER and sentence error rate (SER) were found to be 2.5 % and 19.7 % respectively for the testing set. This WER is significantly lower than the average worker WER of 13.83 %, indicating that the combined transcripts are much more close to the LDC gold-standard transcripts than the individual ones.

4.2. ROVER performance with reliability scores

Next, we incorporated the three reliability scores while combining transcripts using ROVER. Upon tuning the parameters α, β_1 and β_2 on the development set of LDC transcripts, the optimal values were found to be $\alpha = 0.9$, $\beta_1 = 0.6$ and $\beta_2 = 0.2$. It must be noted that the numerical values of these parameters don't give a clear indication of the relative importance of each score, since the force alignment-based score, and hence the overall reliability score, are not normalized to the $[0, 1]$ range. The WER and SER of the reliability-weighted combined transcripts with respect to the LDC gold-standard were found to be 2.3 % and 18.9 % respectively, as shown in Table 1. This represents improvements of 8.0 % and 4.1 % in the WER and SER as compared to the baseline. The fact that we

get appreciable improvement in spite of a very good baseline (2.5 % WER and 19.7 % SER) suggests that the incorporated evaluator reliability metrics do benefit the combination process. This relative improvement in WER upon incorporating reliability information is comparable to the improvement obtained in [7], where the author was combining outputs from multiple ASR systems using ROVER.

Looking at the substitutions occurring in the 2.3 % WER with the LDC reference, we notice that they are mostly due to different transcription conventions between the layman (MTurk worker) and the expert (LDC transcriber). For instance, some of the top errors are: "EX PRESIDENTE" vs "EXPRESIDENTE", "AH" vs"A", "C:O:C:O:P:A" vs "COCOPA" (LDC vs MTurk). The question arises as to whether these are really errors or convention differences. This will be investigated in future work through human verification of the disagreements, and by evaluating the quality of product spoken language components (e.g. WER of an ASR built on the LDC vs MTurk transcripts).

The next step we wanted to consider is whether it is possible to tune the parameters $(\alpha, \beta_1, \beta_2)$ in a completely unsupervised fashion, by using the output of the weighted ROVER as a reference in place of the LDC transcriptions. We found out that the error increases as opposed to using the LDC reference (2.4 % WER, 19.1 % SER), but is still better than the unweighted ROVER performance, in spite of being completely unsupervised. This suggests the possibility of generating high quality transcriptions from multiple non-expert transcriptions without using a gold-standard reference at all (not even for tuning parameters), and will be investigated in the future.

Tuning set →	WER (LDC)	SER (LDC)
Unweighted ROVER	2.5	19.7
Reliability-weighted ROVER	2.3	18.9
% relative improvement	**8.0**	**4.1**

Table 1. WER and SER for the testing set in case of baseline and reliability-weighted ROVER. The LDC transcripts were used for tuning the parameters of the reliability-weighted ROVER.

5. CONCLUSIONS AND FUTURE WORK

This paper presented three metrics for evaluating reliability of transcriptions from multiple non-experts which can be computed without using a gold-standard transcription. These reliability metrics were then used within the ROVER algorithm to combine multiple transcriptions of audio clips collected through the crowd-sourcing website, Amazon Mechanical Turk. Incorporating this reliability information improves the WER and SER of the combined transcripts by 8.0 and 4.1 % respectively (using the LDC transcripts as gold-standard), over a baseline which simply uses word frequency for combination. We also note that there is a difference between the transcription guidelines adopted by LDC, and those used by un-trained workers in Mechanical Turk. Thus, using the LDC transcripts as gold-standard is not fully justified. Furthermore, we investigated the scenario where tuning of all parameters is done without any gold standard transcription, and showed that the benefits are still significant.

We believe that the proposed reliability metrics will become even more important in the case of more challenging acoustic conditions. Our future work will explore more features for capturing transcription reliability, utilize larger databases, and evaluate the performance of speech recognition and machine translation systems trained on such crowd-sourced transcripts.

6. REFERENCES

[1] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks," in *Proceedings of EMNLP*. ACM, 2008, vol. 1, pp. 254–263.

[2] C. Callison-Burch, "Fast, cheap and creative: Evaluating translation quality using Amazon's Mechanical Turk," in *Proceedings of EMNLP*. ACM, 2009, vol. 1, pp. 286–295.

[3] B. Lambert, R. Singh, and B. Raj, "Creating linguistic plausibility dataset with non-expert annotators," in *Proceedings of InterSpeech*. ISCA, 2010, pp. 1906–1909.

[4] M. Denkowski, H. Al-Haj, and A. Lavie, "Turker-assisted paraphrasing for English-Arabic machine translation," in *Proceedings of HLT*. NAACL, 2010, pp. 66–70.

[5] V. Ambati and S. Vogel, "Can crowds build parallel corpora for machine translation systems?," in *Proceedings of HLT*. NAACL, 2010, pp. 62–65.

[6] M. Marge, Banerjee S., and A. I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proceedings of ICASSP*. IEEE, 2010.

[7] J. Fiscus, "A post processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings of ASRU*. IEEE, 1997, pp. 347–354.

[8] B. C. Roy, S. Vasoughi, and D. Roy, "Automatic estimation of transcription accuracy and difficulty," in *Proceedings of InterSpeech*. ISCA, 2010, pp. 1902–1905.

[9] HUB4-NE, "1997 Spanish Broadcast News Speech," *Linguistic Data Consortium*, 1998.

[10] Carnegie Mellon University, "Sphinx-3," *Pittsburgh, Pennsylvania*.

[11] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of ICSLP*. ISCA, 2002, pp. 901–904.

[12] SCTK Version 1.2c, "NIST Speech Recognition Scoring Toolkit," *NIST*, 2008.