Automatic Human Utility Evaluation of ASR Systems: Does WER Really Predict Performance?

Benoit Favre[1], *Kyla Cheung*[2], *Siavash Kazemian*[3], *Adam Lee*[4], *Yang Liu*[5], *Cosmin Munteanu*[3], *Ani Nenkova*[6], *Dennis Ochei*[7], *Gerald Penn*[3], *Stephen Tratz*[8], *Clare Voss*[8], *Frauke Zeller*[9]

[1]Aix-Marseille Univ., [2]Columbia Univ., [3]Univ. of Toronto, [4]City Univ. of New York, [5]Univ. of Texas at Dallas, [6]Univ. of Pennsylvania, [7]Duke Univ., [8]Army Research Laboratory, [9]Univ. College London

Abstract

We propose an alternative evaluation metric to Word Error Rate (WER) for the decision audit task of meeting recordings, which exemplifies how to evaluate speech recognition within a legitimate application context. Using machine learning on an initial seed of human-subject experimental data, our alternative metric handily outperforms WER, which correlates very poorly with human subjects' success in finding decisions given ASR transcripts with a range of WERs.

Index Terms: Automatic speech recognition, Evaluation, User study, Ecological Validity

1. Introduction

Sixty years ago, it was not at all uncommon to think of speech transcription as a worthwhile task in its own right. In part because automated speech transcription works so much better now than then, and in part because it is now so much easier to find, play and store digitized audio, it is no longer the case that transcripts are created purely for their own sake. They are widely used inside spoken utterance retrieval systems, speech summarizers, and speech-to-speech translation systems, where human users never see the transcripts themselves. Actually reading a transcript, however, is still not as easy as reading written text for a variety of reasons. With the exception of accessibility for the hearing impaired, transcription as a speech processing task has been commoditized.

But our means of evaluating the quality of a speech recognition system has remained largely unchanged. Word Error Rate (WER) is still the standard, defined as the number of inserted, substituted or deleted words in the ASR output compared to a reference transcript, divided by the length of the reference, and easily computed by performing a Levenshtein alignment of the two word sequences. It is clear why this is important for evaluating the quality of a transcript, but its applicability to downstream uses of ASR has rightfully come under some scrutiny, e.g., in information retrieval [1] and spoken language understanding [2]. In particular, what is the harm in transcribing some words wrongly if the user never sees the transcript and the performance of the task remains uchanged? In certain domains, at least, relatively high-WER transcripts have been shown to be perfectly usable [3].

The purpose of this paper is *not* to postulate a better alternative to WER for evaluating transcript quality; we stipulate that no better alternative likely exists if the task at hand is taken to be speech transcription for its own sake. Instead, we assert that there are far more useful and relevant tasks to be evaluating speech recognition within than transcription, and, furthermore, that the evaluation of ASR metrics, just like the application task itself, should adhere to the principle of *ecological validity*, which measures how well a real-world situation has been approximated. In our setting, an ecologically valid task is one that realistically simulates how ASR output would actually be used. An ecologically valid *metric* uses that task to predict the performance of subjects using ASR output, i.e. how useful that output would be to humans.

Our contribution is as follows: availing ourselves of the decision audit task as an ecologically valid one, we define the workflow for a metric that involves human-subject experimentation and which has correlates in many other ecologically valid (Section 2) tasks. Then, in order to reduce the cost of running the human experiments, we study how well WER can predict task success, and propose an alternative metric (Section 3). This automatic metric relies on a classifier trained on a number of features of the input transcript and audio (Section 4). Experimental results show that it outperforms WER at predicting human performance (Section 5). Related work is discussed in Section 6.

2. Task

Our decision audit task is similar to [4]. Each study participant[1] (the meeting auditor) plays the role of a recently hired executive product manager in a company that manufactures remote controls. The company had asked three independent teams to design the remote control. The auditor needs to catch up with the decisions made by each design team concerning the remote control in meetings that were held and recorded before s/he was hired. The auditor then needs to browse through the recorded meetings (minutes from the meetings are not available). This description of a real-world scenario helped explain the decision audit task to our participants in meaningful and familiar terms. Hence our participants had a more uniform understanding of what they had to do.

To browse the meetings, the auditors used the JFerret system [5] with a custom user interface. We simplified the interface to show the auditor both the recorded video of the meeting and an extractive summary of the meeting, presented as a clickable list of transcribed utterances. The auditor was able to navigate

This work is supported by NSF award IIS-0845484 and by the Johns Hopkins University Human Language Technology Center of Excellence.

[1]118 subjects participated to this study, including pilots. Results presented in the paper are based on 98 subjects.

through the meeting by clicking on the utterances in the extractive summary. The interface also included the standard play, pause, and stop buttons for the video.

To complete the experiment, our auditors had to operate a desktop computer, fill out a form presented to them on the computer, learn how to operate the JFerret interface during the training portion of the experiment, and become familiar with extracting specific information related to a question that they see on the screen. All of our participants were quite familiar with these subtasks. Perhaps the most challenging of them was extracting information from the meeting data. This is quite similar to taking notes during a class or in meetings with a particular goal in mind, such as passing a course. All of our participants were either students who were already taking notes in class in preparation for assignments and tests, or staff who worked in professional office environments. The fact that our participants were familiar with these subtasks supports the ecological validity of our experimental conditions.

The meetings used in this study come from the AMI corpus's scenario meetings [6]. In each scenario, four participants discuss the design of a remote control over four meetings: a kickoff meeting, design requirements meeting, conceptual design meeting, and detailed design meeting. Our auditors, playing the role of executive product managers with hectic schedules, were given only 25 minutes to browse the final meeting held by each team, which lasted 32–48 minutes.

The extractive summaries contained transcribed utterances from one of five transcription conditions: the reference transcripts as well as four automatic transcripts with WERs averaged over all meetings of 26.8 (ASR1), 28.2 (ASR2), 49.2 (ASR3), and 38.9 (ASR4). The first two were chosen to have nearly identical WERs in spite of having very different acoustic and language models. If WER is an accurate predictor of human-subject performance, then the scores between these two systems should be similar. In a Latin Square experimental design, each participant was to observe three meetings, one with a summary that used a reference transcript and two from summaries that used different automatic transcripts. These conditions thus focussed on the ability of the experimental ASR systems to generate transcripts from which usable summaries could be obtained.

The auditors were instructed to use an adjacent desktop computer to note down decisions related to functionality, physical properties, components, and design. Also, where possible, they had to record any arguments made in the meeting favoring or opposing each decision. To evaluate auditor performance, two judges independently extracted the decisions and arguments by listening to the meetings and viewing the reference transcripts with no time limits. Afterwards, they adjudicated their lists to form the final decision rubrics.

An independent team of markers (who were not judges, experimenters or auditors) used these rubrics to mark each report created by the auditors. Each audit report was marked independently by two markers who then adjudicated their marking reports to come up with the final score. Auditors were assessed according to the number of design decisions they could find, α_1, or partially find, α_2, the number of positive or negative arguments pertaining to the decisions that were found, β, and the number of false alarm decisions, γ that were in fact not made. The score of a subject is then defined as:

$$\text{H-score} = 2\alpha_1 + \alpha_2 + \beta - 2\gamma$$

3. Automatic evaluation metric

Running a human-subject experiment such as this one is time-consuming and expensive. Our objective is therefore to find an automated means of anticipating the results of running a new human-subject experiment, given a new ASR system.

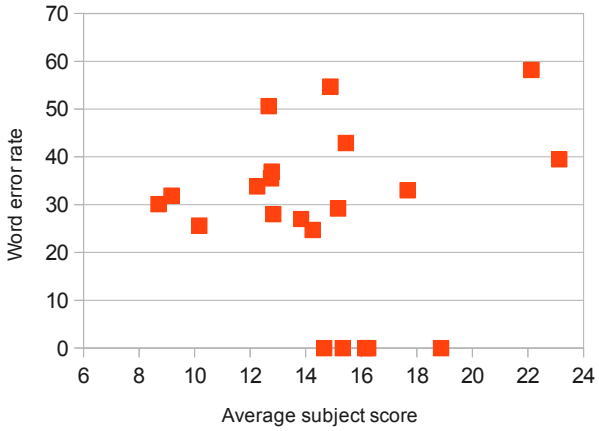

Figure 1: Meeting-level word error rate vs average H-score for all transcript conditions.

The *de facto* automatic metric used in the speech recognition community is word error rate (WER), but our repeated measures ANOVA tests failed to demonstrate any statistically significant effect by WER on audit scores. Figure 1 shows that WER does not numerically correlate with H-score ($\rho = 0.017$), a result in line with those of similar studies [2, 7]. The lack of statistical correlation suggests that, at least for this representative sample, WER is in fact not a good predictor of human performance. There are almost certainly other sources of variability than the quality of the transcript, such as subjects' abilities to find decisions and their mining strategies (listening vs. reading). Reference transcripts do not always result in better H-score than ASR systems because transcript quality is not the only reason for task success, even though it has a major impact. A potential explanation for the lack of correlation is that WER does not capture features of the transcript that condition these other sources.

All the meetings used in the experiment were conducted under the same protocol, had identical goals, and were selected by an experimental design specialist to be as similar to each other according to criteria such as length, number of decisions, flow of dialog, etc.

While WER has been popular and successful at evaluating transcript quality for the sake of measuring how close it is to a gold standard reference, it is not releveant anymore when a human-performed task is considered, which should always be the case for applied research. In the place of WER, we utilize the Auditor Performance Prediction task (APP) which simulates the evaluation for a new ASR transcript. In particular, for each decision-bearing dialog act, the idea is to extract a number of features and train a binary classifier to predict whether the subject found the related decision or not.

4. Features

In order to better simulate human-subject performance, we treat meetings as a discrete-time sequence of dialog acts, each of which is represented by feature values calcuated to capture: the

difficulty of the task (task-specific features), corruption of the transcript (word-error-rate variants and language modeling features), and the importance of the dialog act (linguistic features, such as features that can aid in detecting decisions, and features typically used by utterance extraction systems).

Let w denote the sequence of words in the transcript of a dialog act, w_t be the subsequence of topic words [8], w_s be the subsequence of stopwords, and w_p be the subsequence of non-stopwords. Let T be the constituent parse tree generated by the Berkeley parser [9] over the transcript of the dialog act, and $\mathcal{T}^n = \{t \in subtree(T), |t| \leq n\}$ be the set of all subtrees of T up to size n. Finally, let G be the graph formed from T and w by linking each word to its immediate neighbors in w and to its part of speech in T. In the following, $P(A) \sim \prod_{a \in A} P(a)$ denotes the maximum likelihood estimate of the distribution of the assumed-independent elements of A.

The extracted features are then:

- Task features: the identifier of the user, of the meeting and of the decision; whether the dialog act is shown in the extractive summary.
- Word error rate variants: the dialog-act-level word error rates computed separately on w, w_t, w_p, w_s relative to the reference transcript in the same word space; the value of the tree kernel between T_{ref} and T_{asr} [10]; the Jensen-Shannon divergence (JSD) between $P(\mathcal{T}^n_{asr})$ and $P(\mathcal{T}^n_{ref})$; the graph edit distance[11] between G_{asr} and G_{ref}.
- Language modeling features: the ratio of novel substructures compared to a corpus[2], $|\mathcal{T}^n_{asr}/\mathcal{T}^n_{corpus}|$; the JSD between $P(\mathcal{T}^n_{asr})$ and $P(\mathcal{T}^n_{corpus})$.
- Decision detection features: the number of phrases in T_{asr}; number of pronouns in w; the ratio of the number of 3rd-person pronouns to the number of verbs (indicating a probable use of 1st or 2nd person pronouns) in w; the depth of T_{asr}; the token to type ratio; term frequencies of a list of decision-making verbs and modals ("think", "believe", "should," etc.).
- Summarizer features: the score and rank of the dialog act according to the following summarization baselines: Maximal Marginal Relevance (MMR) [12], KL-divergence, character length, number of words, duration, sum of inverse document frequency, cosine similarity to the centroid of the meeting, topic words, and an acoustic summarization SVM trained on energy, word- and character-normalized duration and pitch.

5. Results and discussion

Given the set of features extracted at the dialog act level, the task is then to train a classifier to predict whether each decision dialog-act was found by the auditor or not. We chose to use an Adaboost classifier that iteratively searches for the best combination of 1,000 decision stumps (one-level decision trees) [13]. This classifier has proved useful in a range of tasks and has the advantage of not being affected by irrelevant features (contrary to SVMs, for instance). As we are interested in predicting human behavior given a new ASR transcript, we performed leave-one-out cross validation, in which, for each transcript condition, a model is trained on the remaining conditions and evaluated on the left-out transcript.

[2] Automated parses from Switchboard and reference parses from the Ontonotes corpora.

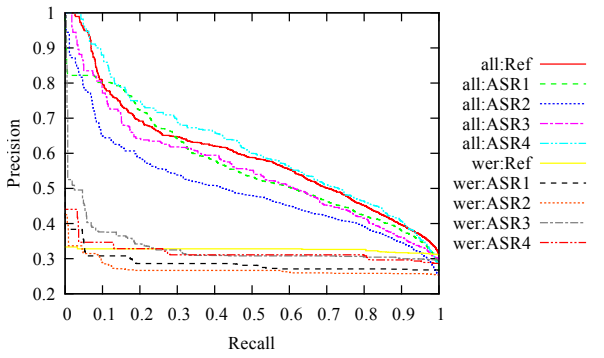

Figure 2: Precision/recall curve for each of the leave-one-out models. The upper family of curves was trained with all features whereas only WER was used in the lower curves.

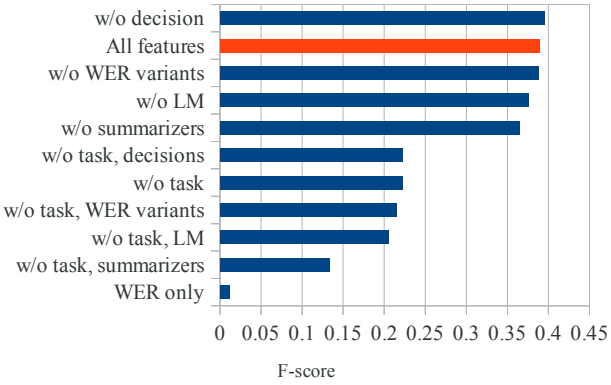

Figure 3: Feature ablation experiment: F-score when each subset of features is removed from training

Figure 2 contains precision-recall curves for the models averaged over each ASR system and overall. There can be little doubt upon observing this figure that features other than WER do indeed improve auditor performance prediction.

Figure 3 shows the impact on the automatic metric quality of removing different subsets of features from training. The most important subset is the set of *task features,* which capture that a particular decision is difficult to detect from a dialog act, or that a particular auditor is not good at finding decisions. The *summarizer features,* which capture whether a dialog act is important for the meeting in general, seem to help in predicting auditor success. Interestingly, *decision detection features* did not improve performance. This might be because ASR errors disrupt the Berkeley parser, which make those features less reliable. In any case, many features are affected by the quality of the transcript, including all of the parsing-related features. On the other hand, features like the length summary baseline are more independent of the transcript and can only serve as normalizers for the other features. Notice that, in comparison to any one of these ablated classifiers, WER is still very poor at predicting human-subject performance. WER variants are also of little importance, given the ablation experiments.

6. Related work

Several previous studies have lamented the lack of tight alignment or correlation between WER and downstream task perfor-

mance, such as spoken information retrieval, spoken language understanding, and spoken machine translation. Sanders and Le [14] studied the effects of speech recognition accuracy on dialog systems, and found a high correlation between WER and task completion — even for high WERs. Even here, WER was shown to have little effect on user satisfaction when it is less than 35%), however.

There have also been previous attempts at modifying WER to take into account different types of errors in ASR output, e.g., using an error rate that weighs content words or information bearing words more heavily [1]. Morris et al. [15] proposed to use match error rate (MER) and word information loss (WIL) to evaluate recognition performance and represent the proportion of word information communicated. For high error rates, they found that these are more appropriate. Similarly McCowan et al. [7] suggested posing recognition evaluation as an information retrieval problem and thus using a more application-oriented evaluation. Mishra et al. [16] developed a metric called Human Perceived Accuracy, and showed in a voice mail recognition task that it correlates more highly with human judgments of ASR accuracy than WER does. Their method of using a regression task for predicting recognition performance has some similarity to our study in this paper, although our focus is on the metric of ASR in the context of an ecologically valid task. This is important — without ecological validity, the human judgements and scores themselves are meaningless. Human judgements of ASR accuracy are not ecologically valid.

Current research on MT evaluation has many parallels to the issues facing ASR evaluation. In MT, automated metrics are routinely used in spite of their now well-documented limitations [17] because they, too, provide a rapid, cost-effective means for developers to tune their systems' performance. To independently validate proposed metrics, NIST Metrics MATR [18] and WMT organizers [19] have conducted shared meta-evaluation tasks alongside the standard MT evalation tasks. Initially these tested for correlation with human-subject judgments of translation adequacy and fluency — another ecologically invalid pair of tasks that cannot address real-world scenarios where MT provides support to downstream tasks that humans actually perform. Most recently, they have introduced a "quality estimation" (QE) task, which evaluates MT quality in terms of its impact on human post-editors, just as in our approach anchors its evaluation of ASR in a downstream decision-audit task. The majority of the QE system developers also make use of parsers, part-of-speech taggers, named entity recognizers, etc. to derive linguistic features of the source and target language texts, and then train M5P regression trees or SVM regression models on different combinations of these features to estimate the level of human effort required to post-edit MT output.

7. Conclusion

WER has a legitimate place in the evaluation of speech recognition systems, but predicting human-subject performance on realistic applications of ASR may not be one of them. Complementary evaluation measures to WER are necessary in order to determine the effects of a change to an ASR system in an ecologically valid context.

Human-subject judgements are expensive to collect, but automatically learning to predict those judgements, as in the dialog act classifier here, brings such complementary measures closer to the grasp of experimenters who wish to regression-test their ASR improvements on real applications.

A significant remaining problem is the portability of judgements collected on one task to another, which has not been addressed here.

8. References

[1] J. Garofolo, C. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*, 1999.

[2] Y. Wang, A. Acera, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2003, pp. 577–582.

[3] G. P. C. Munteanu, R. Baecker, E. Toms, and D. James, "The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives," in *in Proc. of CHI*, 2006, pp. 493–502.

[4] G. Murray, T. Kleinbauer, P. Poller, T. Becker, S. Renals, and J. Kilgour, "Extrinsic summarization evaluation: A decision audit task," *ACM Transactions on Speech and Language Processing*, vol. 6, no. 2, pp. 1–29, 2009. [Online]. Available: http://doi.acm.org/10.1145/1596517.1596518

[5] "Augmented multiparty interaction," "http://www.amiproject.org/showcase/meeting-browsers", May 2012.

[6] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," in *In: Proc. Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.

[7] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," IDIAP (Institut Dalle Molle dIntelligence Artificielle Perceptive), Martigny, Switzerland, Tech. Rep. IDIAP-RR 04-73, March 2005.

[8] C. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2000, pp. 495–501.

[9] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning accurate, compact, and interpretable tree annotation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 433–440.

[10] A. Moschitti, "Making tree kernels practical for natural language learning," in *Proceedings of EACL*, vol. 6, 2006, pp. 113–120.

[11] D. Justice and A. Hero, "A binary linear programming formulation of the graph edit distance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 8, pp. 1200–1214, 2006.

[12] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing

summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336.

[13] B. Favre, D. Hakkani, and S. Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[14] G. A. Sanders, A. N. Le, and J. S. Garofolo, "Effects of word error rate in the darpa communicator data during 2000 and 2001," in *Proc. of ICSLP*, 2002.

[15] A. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in *Proceedings of Int. Conf. Spoken Language Processing (INTERSPEECH-ICSLP-04)*, 2004.

[16] T. Mishra, A. Ljolje, and M. Gilbert, "Predicting human perceived accuracy of asr systems," in *Proc. of Interspeech*, 2011.

[17] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of bleu in machine translation research," in *Proc. of EACL*, 2006, pp. 249–256. [Online]. Available: http://aclweb.org/anthology-new/E/E06/E06-1032

[18] K. Peterson, M. Przybocki, B. Antonishek, M. Yilmaz, and M. Michelf, "Metricsmatr10: Evaluation overview & summary of results," Presentation at WMT10 & NIST MetricsMaTr10, ACL 2010.

[19] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2012 workshop on statistical machine translation," in *Proc. of 7th SMT Workshop*. ACL, 2012.