END-TO-END SPEECH RECOGNITION ACCURACY METRIC FOR VOICE-SEARCH TASKS

Michael Levit, Shuangyu Chang, Bruce Buntschuh, Nick Kibre

Speech at Microsoft, Microsoft Corporation

ABSTRACT

We introduce a novel metric for speech recognition success in voice search tasks, designed to reflect the impact of speech recognition errors on user's overall experience with the system. The computation of the metric is seeded using intuitive labels from human subjects and subsequently automated by replacing human annotations with a machine learning algorithm. The results show that search-based recognition accuracy is significantly higher than accuracy based on sentence error rate computation, and that the automated system is very successful in replicating human judgments regarding search quality results.

Index Terms— voice search, semantic accuracy

1. INTRODUCTION

There is a well established set of success criteria for ASR. Among them, word error rate (WER) and sentence error rate (SER) are the two most commonly used. However, these metrics (as well as others such as perplexity of language models) often fail to reflect the real usefulness of the speech recognition system to the user. Except for the dictation task, where speech recognition is the means and purpose of the task, most other systems avail themselves of ASR to achieve higher level goals, for instance, execute a spoken command or connect a call to the right operator. As a result, perfect speech recognition and ultimate task success cannot be considered interchangeable quantities any longer. A command can be recognized but not understood; similarly recognized text does not have to be exactly what the user has said to turn into a desirable meaningful interpretation.

Because of this, a number of alternative task-specific metrics have found their applications across the field. Spoken dialog systems realized via context-free grammars [1] (e.g. early call routers) typically reduce the language of the encoded word sequences to an enumerable set of semantic values. Open-ended dialog systems allow for practically unconstrained language but end up internally representing it as a finite hierarchy of pre-defined topics (e.g. call types) possibly accompanied by a number of topic specific parameters [2, 3]. Another example is the directory assistance systems that are successful as long as they correctly recognize the locality and business names parts of the requests.

Voice search applications demand their own task specific success criteria, and it is clear that these criteria must have to do with users' objective of finding relevant search results for their queries. Similarly to the examples above, even if automatic transcriptions produced by ASR contain errors, the quality of the search results found by the downstream search engine is not doomed to degrade. For instance, misrecognizing *"the home depot"* as *"home depot"* has no effect on the found web pages. Similarly, misrecognition of *"another one bites the dust"* as *"another one bites the dusk"* should not stop any search engine from finding the right sites.

On the other hand, it is well known that human transcriptions, though expected to deliver a gold standard for speech transliteration, are also prone to errors or controversial decisions. However, many of these errors are of pure lexical nature and do not affect semantics. For instance, a transcriber could spell *"Facebook"* as *"face book"*. Many other transcription errors are mere typos that do not affect our understanding either (e.g. *"Gooogle"*).

Good evaluation methods need to be able to disregard these irrelevant misrecognitions and only focus on the parts that matter. We started off by experimenting with various normalization options trying to address lexical variations deemed semantically irrelevant. For instance, instead of computing SER, we would remove spaces from both references and hypotheses, and then see whether the two obtained strings are equal. Other normalization options included removing apostrophes and even ignoring plurals. While this normalization helped ignoring insignificant differences in some cases, it failed in many others. The above case of *"the home depot"* is one example. On the contrary, recognizing *"you r good"* instead of *"your good"* or *"who"* instead of *"the who"* leads to clearly bogus search results.

Another common approach to address the inadequacy of metrics such as WER is to give greater weight to information bearing words. In [4, 5] alternative error metrics were proposed that incorporated word salience into the computation. For the document retrieval task, the metric in [4] relied on instances of named entities, stop words, and salient query words, and exhibited better correlation with the retrieval accuracy. The metric from [5] assigned IDF-based salience weighting to individual words. Our preliminary experiments indicated that even these approaches would still not be able to solve many of the issues in voice search.

Thus it became obvious that one does indeed need to include search engine in the ASR evaluation loop. One exam-

ple of how this can be done is the Web Score, described in [6] where speech recognition is considered a success when the top search results for reference and hypothesis are identical. However, our experiments indicated that this metric can be insufficient, especially in cases where there is no obvious best result or where there are several of them. In the remainder of this paper we describe our approach to a search-based speech recognition accuracy metric that is based on human intuition. First, in Section 2 we explain the motivation behind it and describe the process of obtaining gold standard references for search-based ASR evaluation. Then, in Section 3 a regression method for replicating these references will be described. Section 4 is dedicated to analysis of the new metric and how it compares to other success criteria. We complete the paper with a discussion section.

2. INTUITIVE ERROR METRIC FOR SEARCH RESULTS

In search engine tuning, it is common to rely on the discounted cumulative gain (DCG) and its normalized variants. DCG is essentially an average of human graded relevance scores of search results, weighted by the position of each result [7]. Similarly, for the task of assessing impact of speech recognition errors on the end-to-end user experience, we decided to start by relying on human intuition. Our application domain is Bing Voice Search for Mobile where users run a smart phone application to type or speak search queries that will be passed to Bing search engine, with search results displayed on screen. We compiled a training set of 3K voice queries to have roughly equal number of correctly and incorrectly recognized utterances (i.e. 50% SER).

Instead of using DCG directly, for each voice search query in our training set, we took its manual transcription (*reference*) and recognition output (*hypothesis*) and sent both to the Bing search engine retrieving up to 20 top-ranked search results for each. We would then present the two sets of results to two human labelers, asking them to grade the quality of the search results for the hypothesis given the search results for the reference. In order to solicit intuitive judgments, we provided the labelers with only a minimum amount of instructions, asking them to issue one of the four possible grades:

0: search results are very different
1: search results exhibit some similarity
2: search results are very similar
3: search results are identical (or nearly identical)

If search results for the reference were of insufficient quality, we still recommended the labelers to make judgments solely based on search result comparison.

After the first 100 queries, we let the labelers discuss their annotations with each other to facilitate calibration. After that, they proceeded to annotate the remaining queries[1]. Table 1 shows the confusion matrix for the two labelers annotations on utterances with different reference and hypothesis texts. The labelers agreed on 85% of all queries, with the inter-labeler $\kappa = 0.7$. However, merging labels 0 and 1, and 2 and 3 respectively, resulted in $\kappa = 0.86$. Agreement was the easiest to achieve on the ends of the spectrum (that were also best represented). Most of the disagreements could be attributed to only a 1-grade difference (0 vs. 1, 1 vs. 2, or 2 vs. 3), and in these cases we took the mean as the final label. Only 36 cases were due to a difference by two or more grades, and the labelers were asked to adjudicate them.

labeler1 \ labeler2	0	1	2	3
0	956	74	5	3
1	16	41	9	3
2	10	38	43	7
3	3	12	53	248

Table 1. *Inter-labeler agreement search-based scores.*

2.1. Using Annotations to Judge Recognition Accuracy

We then introduced the obtained labels in the recognition evaluation framework. First, we linearly normalized the annotator-provided grades to the $[0; 1]$ range (grade 3 being mapped to 1.0) and denoted the obtained values *"search quality score"* (SQS). Let there be N queries in the test corpus. Assuming that each recognition is accompanied by its confidence, and that recognition result will be accepted iff its confidence is greater than some pre-specified confidence threshold θ, we define two corpus-level metrics as functions of θ: correct accept rate (CA) and false accept rate (FA).

$$\mathrm{CA} := \frac{1}{N} \times \sum_{i=1}^{N} \begin{cases} \mathrm{SQS}_i & \text{if } \mathrm{conf}_i \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\mathrm{FA} := \frac{1}{N} \times \sum_{i=1}^{N} \begin{cases} 1 - \mathrm{SQS}_i & \text{if } \mathrm{conf}_i \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

We can then plot a *"Voice Search Quality"* (VSQ) curve as a dependency of CA on FA subject to changing θ. In Section 4 we will show how this evaluation criterion compares to the more traditional sentence error rate (SER) based curves in which CA and FA are also computed according to (1) and (2) except that a binary match metric is used instead of (generally real-valued) SQS. We will also note the differences between reference and hypothesis word strings for which VSQ curves show different sensitivity than SER.

3. REPLICATING SEARCH SCORES

Having humans evaluate similarities of sets of search results is not only very expensive but also time consuming. For instance, our two annotators spent between 30 seconds and 2 minutes per query. Obviously, this is not acceptable for an evaluation procedure that is supposed to be run every time there is a change in the test set or in the system. Therefore, our next goal is to remove humans from this annotation loop[2].

[1] In reality, only half of the queries required manual annotations for references and hypotheses were identical in the rest of the cases.

[2] We would still use human annotators to produce utterance transcriptions.

To do that, we decided to extract a plurality of observable features from the pairs of search results sets and use statistical learning to map these features onto the space of SQS scores.

3.1. Classification Features

There are several groups of features that we included in our experiments:

- absolute numbers of search results for reference and hypothesis
- can reference and hypothesis be reduced to one another using Bing spell checking mechanisms
- weighted recall and precision of found page titles on both lists with weights reflecting source and target ranks. For instance, the exact formula for recall is:

$$R := \frac{1}{Z} \times \sum_{i=1}^{N} \frac{1}{i} \times \frac{N - \max(H_i - i, 0)}{N} \quad (3)$$

  where N is the maximum number of search results to consider, i iterates over ranks of all reference search results, H_i is the rank of the i^{th} reference search result among the search results found for the hypothesis (and is set equal to $N + i$ if hypothesis search results do not contain it), and Z is a normalization coefficient $\sum_i 1/i$.
- similar metrics for word 1/2/3-grams in the titles
- we also took advantage of the RANKER tool developed by the BING search team to automatically evaluate appropriateness of an arbitrary page for an arbitrary query. The tool is trained to produce scores comparable across query/page pairs. Specifically, we compute four measurements: average score of hypothesis search results for the hypothesis query (weighted by page ranks), average score of reference search results for the hypothesis query, as well as analogous two metrics for the reference query.

Thus, a total of 15 features were extracted for each of the 1500 queries for which ASR misrecognized at least one word.

3.2. Results

We trained a linear SVM regression model to approximate SQS values via features above. To compensate for a small sample size (consequence of a laborious labeling process), 5-fold cross-validation was employed. With target values all lying between 0 and 1 (cf. Section 2.1; average of 0.26; standard deviation of 0.39) the mean square error of the regressed values was measured to be 0.038. Alternatively, casting the problem as classification (with four classes corresponding to the original categories suggested to labelers) and using boosting to compute classification accuracy, we obtained classification accuracy of about 90%. Most confusions happened between neighbor classes. To put this numbers in perspective, we plotted two VSQ curves: one for SQS derived from manual annotations, and another from predicted SQS values. In Figure 1 these two curves appear virtually identical for all confidence thresholds, meaning that we have successfully excluded humans from the annotation loop and can conduct our further analysis based on the predicted SQS values. It

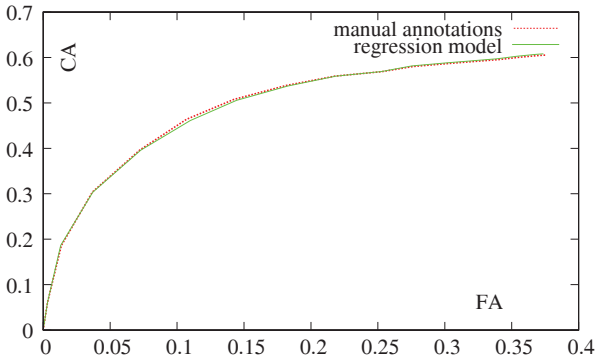

Fig. 1. *VSQ curves based on regressed SQS values replicate nearly perfectly the curves based on manual labels.*

should be noticed that the features from the ranking group contributed to classification accuracy about 1.5% absolute. However, since obtaining these features is time consuming, in some practical settings we would also experiment without them. Fortunately, it appeared that absence of these features did not introduce any bias in the VSQ metric.

4. VSQ CURVES VS TRADITIONAL ERROR METRICS

In this section we compare voice search quality curves with traditional criteria used to evaluate speech recognition quality, such as sentence error rates and variations thereof.

There are two dimensions to direct our investigations along. First, we provide a quantitative analysis of the numbers, and then dive into individual cases that explain the differences in corpus-based metrics.

For 3K voice search queries, Figure 2 plots the VSQ curve (against regressed labels) and three baselines:

1. SER-based curve: each utterance contributing either 1.0 (if reference and hypothesis are the same) or 0.0;
2. similar but both reference and hypothesis are normalized to remove spaces, apostrophes etc.
3. search-based evaluation approach from [6]: recognition is considered a success iff the first top-ranked search results for reference and hypothesis agree.

We see that the VSQ curves (that we consider a direct derivative from human intuition and, therefore, the most reliable way of estimating misrecognition impact on user experience with the search application) are conveying a different message than other baselines. At zero confidence threshold, the search-based accuracy rates are about 15% absolute higher than measured sentence accuracy, and more than 10% absolute higher than sentence accuracy after text normalization. In fact, the VSQ accuracy is closer to the word accuracy (67%

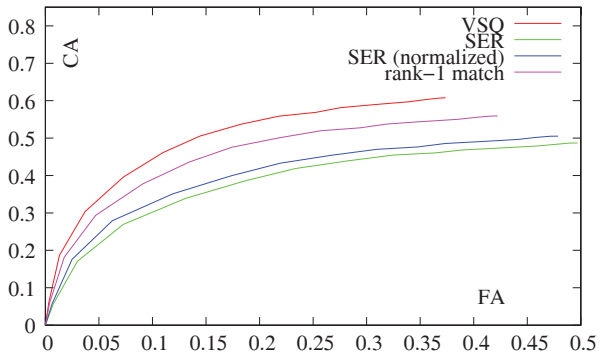

Fig. 2. *VSQ curves reflecting human intuition are quite different from all other baselines.*

in this case) than to sentence accuracy. Moreover, the curves indicate that identity of top-rank search results does not constitute a good estimate of user experience with the system, but underestimates it by a significant margin.

To better understand differences between normalized SER and search-based accuracy, we analyzed the queries whose normalized-SER and search-based evaluations differed. There were 472 such queries, and only for 6 of them the search-based evaluation assigned a lower score than normalized SER. For instance, *"incase"* was recognized as *"in case"*. We therefore focused on the cases where high SQS corresponded to SER-based FA (60% of all cases). Of these (purported) misrecognitions, 15% were due to plural forms, additional 20% were caused by errors in one-letter and/or function words, and further 20% were accounted by synonyms, alternative spelling variants or typos. The rest of the misrecognitions could be ascribed to miscellaneous causes, such as missing content words that did not stop the search engine from finding good hits.

We have experimented with the VSQ in three different setups. First, to make sure the curves remain sensitive to small random language model changes, we pruned our language model with seven different thresholds (Stolcke pruning). The results showed that confidence zero VSQ numbers correlated strongly (Pearson coefficient 0.998) with the sentence accuracy. In the second setup we ran spelling correction on ASR results. While this improved SER by 2% relative, the VSQ curves, as expected, were not affected. Thus, the VSQ criterion proved to be robust against ASR (and human transcription) errors that can be explained by typos and other factors that do not affect understanding. For the final investigation, we have augmented the main SLM with a number of additional LMs specialized on address capture, weather queries and other domains. From the past experience we knew that this improves SER by 2% relative. However, search-based evaluation showed almost no gain. Thus, while domain specific grammars helped the ASR to produce more accurate recognition texts, they did not significantly improve users' overall experience with the search application, a conclusion that helped us re-allocate resources to focus on truly relevant ASR improvements.

5. DISCUSSION

We have presented a novel metric for search-based evaluation of speech recognition systems in a voice search scenario. Unlike traditional figures-of-merit, this metric is motivated by the real user experience, and captures how this experience is affected by various recognition errors. We have shown that true impact of many recognition errors on the search quality is significantly smaller than what measured sentence error rates might suggest, and that the new metric is more robust against transcription errors and "forgivable" ASR errors such as alternative spellings or function words. Several caveats should be kept in mind while using this metric. First, quality of search results is not the only factor that contributes to a good user experience. Displaying correctly recognized text can also be important as some users would not wait for the search engine to return, if they do not like the recognized text. Second, we currently lack an empirical confirmation that comparing search results for cases where these results are not satisfactory even for the reference texts is the right approach. We plan on extending our investigation in both of these directions. Finally, since search engines evolve, it is possible that the same recognition results will be judged differently over time. Nonetheless, this methodology proved to be very beneficial for our ASR. While not improving speech recognition per se, it allowed us to focus on those directions of improvement that are important for the end-user.

6. REFERENCES

[1] W3C, "Semantic Interpretation for Speech Recognition," April 2007, url: http://www.w3.org/TR/semantic-interpretation.

[2] Gorin A., Riccardi G., and Wright J., "How May I Help You?," *SpeechCom*, vol. 23, pp. 113–127, 1997.

[3] Levit, M., *Spoken Language Understanding without Transcriptions in a Call Canter Scenario*, Ph.D. thesis, Erlangen University, Erlangen, Germany, 2005.

[4] Garofolo J. et al., "TREC-7 Spoken Document Retrieval Track Overview and Results," in *Proc. 7th Text Retrieval Conference TREC-7*, 1998.

[5] T. Mishra, Ljolje A., and Gilbert M., "Predicting Human Perceived Accuracy of ASR Systems," in *Proc. Interspeech2011*, 2011.

[6] Chelba C. and Schalkwyk J. et al., "Language Modeling for Automatic Speech Recognition Meets the Web: Google Search by Voice," OGI CSLU Seminar, 2011.

[7] Croft B., Metzler D., and Strohman T., *Search Engines: Information Retrieval in Practice*, Addison Wesley, 2009.