Search-based Evaluation from Truth Transcripts for Voice Search Applications

François Mairesse
Amazon.com
101 Main street, Suite 900
Cambridge, MA 02142, USA
mairessf@amazon.com

Paul Raccuglia
Amazon.com
101 Main street, Suite 900
Cambridge, MA 02142, USA
praccu@gmail.com

Shiv Vitaladevuni
Amazon.com
101 Main street, Suite 900
Cambridge, MA 02142, USA
shivnaga@amazon.com

ABSTRACT

Voice search applications are typically evaluated by comparing the predicted query to a reference human transcript, regardless of the search results returned by the query. While we find that an exact transcript match is highly indicative of user satisfaction, a transcript which does not match the reference still produces satisfactory search results a significant fraction of the time. This paper therefore proposes an evaluation method that compares the search results of the speech recognition hypotheses with the search results produced by a human transcript. Compared with a strict sentence match, a human evaluation shows that search result overlap is a better predictor of (a) user satisfaction and (b) search result click-through. Finally, we propose a model predicting the Expected Search Satisfaction Rate (ESSR), conditioned on search overlap outcomes. On a held out set of 1036 voice search queries, our model predicted an ESSR within 0.9% (relative) of the ground truth satisfaction averaged over 3 human judges.

Keywords

Evaluation, voice search, automatic speech recognition, search engines, entity resolution

1. INTRODUCTION

Voice search applications are typically evaluated by comparing the automated speech recognition (ASR) output to a single human reference transcript, and measuring either the Word Error Rate (WER) or the ratio of exact string matches (Sentence Error Rate or SER) [4, 1]. We find that while an exact sentence match is highly indicative of user satisfaction, a sentence error is a poor indicator of user dissatisfaction (see Sections 3 and 4). Sentence errors are not predictive of user satisfaction because transcribers produce a single transcript per audio sample, while many other transcription variants could produce satisfactory search results. Table 1 shows how a normalization variant between *t-shirts* and *t shirts* affects search results in a product search engine.

SIGIR '16, July 17 - 21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914735

An error analysis reveals that 37% of the SER is due to arbitrary normalization choices (e.g. *tooth brush* vs *toothbrush*), function word deletion (e.g. *a*, *the*), possessives, plurals, or transcription errors. This is partly due to the fact that the language model is trained on search queries, hence the system is likely to output search-like utterances. A first solution is to handcraft normalization rules, however those do not scale and can mask errors that impact search (e.g., ignoring white-space mismatches could artificially increase accuracy). Another solution would be to ask transcribers to provide an exhaustive set of transcripts for each audio sample, however it is not cost effective.

In contrast, information retrieval systems are evaluated by asking human judges to estimate the relevance of individual results, and then compute the Mean Average Precision or Normalized Discounted Cumulative Gain (NDCG) [3]. Such metrics are costly when used for evaluating voice search systems, since the human judgments must be collected for new queries resulting from ASR model updates. While previous work has focused on jointly optimizing ASR and information retrieval systems [2], it assumes the existence of search relevance judgments for every test query. Our goal is to remove the need for collecting large amounts of search evaluation data for offline testing, by building a user satisfaction model from a small set of judgments.

In order to evaluate the effect of voice recognition errors more accurately without requiring additional transcription resources, in Section 2 we propose an evaluation method that compares the search results of the speech recognition hypotheses with the search results produced by a human transcript. This is related to the Google WebScore, which is defined as the number of times the first result of the ASR output matches the first result of the human transcript [5]. However, we find that using search results beyond the first one produces metrics that correlate better with human judgment. Section 3 reports experiments showing that partial overlap metrics outperform SER as a predictor of human search result ratings and user clicks. Finally, Section 4 proposes a model of user satisfaction conditioned on search overlap outcomes. Experiments on a held out test set of 1036 utterances predicted an Expected Search Satisfaction Rate (ESSR) within 0.9% relative to the ground truth satisfaction averaged over 3 human judges.

2. SEARCH-BASED EVALUATION

An ideal search-based evaluation would compare the returned search results with the set of satisfactory results that the user intended to find. Labeling search results is costly

Table 1: Search results for the queries *t-shirts* and *t shirts*. Results that occur in both searches within the first 10 results are in bold.

Search query: *t-shirts*	Search query: *t shirts*
1) **Hanes Men's 4 Pack Short Sleeve Comfortsoft Tee**	1) **Fruit of the Loom Men's 4-Pack Pocket T-Shirt - Colors May Vary**
2) Hanes Men's 5 Pack Comfort Soft Crew Neck Tee, White, Large	2) **Russell Athletic Men's Basic Cotton Tee**
3) **Fruit of the Loom Men's 4-Pack Pocket T-Shirt - Colors May Vary**	3) **Problem Solved T-shirt Funny Shirts**
4) **Russell Athletic Men's Basic Cotton Tee**	4) **Hanes Men's 4 Pack Short Sleeve Comfortsoft Tee**
5) **Problem Solved T-shirt Funny Shirts**	5) Soffe Men's Men'S Long Sleeve Cotton T-Shirt
6) **Fruit of the Loom Men's 4-Pack Crew Neck T-Shirt**	6) Original Penguin Men's Bold Stripe Print Tee
7) Next Level Mens Premium Fitted Short-Sleeve Crew	7) California Republic Men's T-Shirt
8) **Mens Funny Sayings Slogans T Shirts-I May Be Wrong tshirt-Ideal Gift Ideas**	8) **Fruit of the Loom Men's 4-Pack Crew Neck T-Shirt**
9) Hanes Men's Classics 6 Pack Crew Neck Tee	9) Fruit of the Loom Men's Short Sleeve Crew Tee
10) Hanes Women's Relax Fit Jersey V-Neck Tee	10) **Mens Funny Sayings Slogans T Shirts-I May Be Wrong tshirt-Ideal Gift Ideas**

and prone to ambiguities, since the annotator does not know what the user had in mind. However, human transcripts can be used to estimate the intended search results. Let us define $R_{ref}^{1:N}$ as the set of the first N results returned by sending the *reference* transcript to search, and $R_{asr,1}^{1:N}$ the first N results returned by sending the *1-best ASR hypothesis* to search.

Since the true user intent is unknown, we rely on R_{ref} as an approximation to the search results the user is looking for. Table 1 shows that search engines sometime fail to return consistent results between equivalent surface forms. As a result, the next section shows that a strict ordered result match correlates poorly with human judgments of search accuracy ($r = 0.65$), i.e., it over-penalizes the system. This paper therefore investigates a family of *search overlap* metrics characterizing the minimum number of results in common N_{min} between the two sets of N search results:

$$o(R_{asr}, R_{ref}, N_{min}, N) =$$
$$\begin{cases} \text{undefined} & \text{if } |R_{ref}| = 0 \\ 1 & \text{if } |R_{asr,1}^{1:N} \cap R_{ref}^{1:N}| \geq min(N_{min}, |R_{ref}|) \\ 0 & \text{otherwise} \end{cases}$$

Let us abbreviate $o(R_{asr}, R_{ref}, N_{min}, N)$ as $o(N_{min}, N)$. The search results in Table 1 have 6 results in common. Since the first 2 results do not overlap we can measure that $o(1,2) = o(2,2) = 0$. The first 4 results have exactly 3 items in common, hence $o(1,4) = o(2,4) = o(3,4) = 1$ and $o(4,4) = 0$. Note that the Google WebScore corresponds to $o(1,1)$ since it only takes the first result into account [5].

3. CORRELATION WITH HUMAN JUDGMENTS AND USER CLICKS

In order to identify the most suitable metric for evaluating ASR systems, we asked human judges to evaluate 3500 search results produced by the 1-best hypothesis of a mobile shopping voice recognition application, based on the corresponding human reference transcript. See instructions in Figure 1. Human judges associated each ASR hypothesis with a search quality rating, and discarded any non-shopping utterances. Additionally, we asked human judges to evaluate the search results produced by the reference transcript. Note that in a first experiment 2 judges rates 500 voice searches, and in a follow up experiment 6 groups of 3 judges each rated additional slices of 500 voice searches, for a total of 3500 utterances.

1. Press play to listen to the audio
2. Evaluate the search results on a scale from 1-3:
   - 1: I am not satisfied by the search results
   - 2: I am partially satisfied by the search results
   - 3: I am satisfied by the search results
   - N/A: not a shopping query, or I cannot make sense of what is being said after looking at the results page

Figure 1: Search evaluation instructions.

Each voice search with a rating of 3 is labeled as *satisfactory* while ratings below 3 are labeled as *unsatisfactory*. The *Rating* column of Table 2 shows the Pearson's correlations between the satisfaction ratings and the evaluation metrics mentioned in the previous section (for various values of N_{min} and N), as well as sentence error rate (SER). In order to focus on the effect of ASR errors, these were obtained after removing non-shopping utterances (e.g., navigation queries, phatic utterances) and utterances for which the search results of the reference transcript were labeled as unsatisfactory. Out of the remaining 2462 utterances, a training set of 1426 utterances was used in Table 2, while 1036 utterances from distinct groups of annotators were held out for testing.

Results show that whether or not the reference results and the ASR results have one result in common (i.e., *o(1,10)*) is the best predictor of human ratings ($r = .82$). All proposed search-based metrics are better predictors of search ratings than sentence match ($r = .60$). We find that the Google WebScore (i.e., *o(1,1)*) does not predict satisfaction as well as metrics that go beyond the first search result ($r = .71$).

Additionally, we investigated whether overlap metrics can predict user clicks following a voice search. The *Click* column in Table 2 shows the correlations between the result overlap and click events over 63,000 live voice queries. A query is associated with a click event only if the user clicks

Table 2: Pearson's correlation coefficient between overlap and SER metrics with (1) human search rating and (2) whether or not a user clicked on a search result. Largest correlations are in bold.

Metric	Correlation with rating	Correlation with click
o(1,10)	**0.82**	**0.22**
o(1,5)	0.81	0.21
o(1,3)	0.78	0.20
o(3,5)	0.72	0.16
o(1,1)	0.71	0.15
o(10,10)	0.62	0.04
Ordered result match	0.61	0.02
Sentence match	0.60	0.14
Word error rate	0.58	0.16
# utterances	1,426	63,000

on a product on the first page returned by the voice search. The results confirm that low result overlap metrics are better predictors than sentence match ($r = .22$ vs $r = .14$, respectively), however the correlations are weaker. This could be partly due to users not clicking on a result even though the result is relevant. Additionally, the reference search results were retrieved one month after the live results, hence the variation of the search results over time is likely to weaken the correlations too.

4. USER SATISFACTION PREDICTION

This section focuses on estimating a user satisfaction model based on the features investigated in the previous sections. Since the satisfaction model will be used to evaluate production ASR systems, it is important to (1) avoid any overfitting and (2) be able to inspect the learned model. For these reasons, we estimate a discrete conditional probability table based on rating counts of the training data detailed in Section 3. Table 3 shows the estimated probability distribution of user satisfaction conditioned on the search overlap. Given such a table, one can compute the expected probability of satisfaction over a set on unseen utterances by taking the average of the probability that the search results were rated as satisfactory, over all utterances in the test set.

In order to focus the attention on the effect of ASR on search accuracy, Table 3 only includes utterances for which the reference transcript's results are rated as satisfactory. Hence our approach provides an upper bound on user satisfaction. The method could trivially be extended to include utterances for which the search engine did not return relevant results to model the effect of search engine errors.

While observing an overlap is not more indicative than observing a sentence match, the absence of overlap is a better indicator of poor rating than a sentence mismatch ((P(r=non-sat|¬ o(1,10))=.79) vs P(r=non-sat|¬ sentence match)=.46).

This table provide a model of the expected user satisfaction s resulting from an ASR output given that the reference transcript produces valid results, by marginalizing over all search overlap outcomes. Because we focus on voice queries for which the reference transcript is a valid search, the probability of satisfaction is 1 if the ASR output matches the reference transcript. Hence, the overlap metric comes into play only when the transcripts do not match. As a result

Table 3: Probability of satisfactory rating (r=*sat*) conditioned on various metric outcomes, computed over 1426 training utterances. Utterances for which the reference transcript results in a non-satisfactory search were removed.

Rating r	r = non-sat	r = sat
P(r|o(1,10))	0.02	0.98
P(r|¬o(1,10))	**0.79**	**0.21**
P(r|o(1,3))	0.01	0.99
P(r|¬o(1,3))	0.71	0.29
P(r|o(3,5))	0.01	0.99
P(r|¬o(3,5))	0.61	0.39
P(r| sentence match)	0.00	1.00
P(r|¬ sentence match)	**0.46**	**0.54**

Table 4: Probability of satisfactory rating (r=*sat*) given a sentence mismatch, computed over the 486 training utterances not matching the reference transcript.

Rating r	r = non-sat	r = sat
P(r|o(1,10),¬sentence match)	0.08	0.92
P(r|¬o(1,10),¬sentence match)	0.79	0.21
P(r|o(1,3),¬sentence match)	0.06	0.94
P(r|¬o(1,3),¬sentence match)	0.71	0.29
P(r|o(3,5),¬sentence match)	0.05	0.95
P(r|¬o(3,5),¬sentence match)	0.61	0.39

our final model relies on the estimated satisfaction distribution conditioned on a sentence error, as shown in Table 4. The estimates in this table are computed over the subset of 486 training utterances for which the ASR output did not match the reference transcript. This table is our final user satisfaction model.

Equation 1 illustrates how the expected user satisfaction is computed based on the conditional probability table.

$$P(sat) = \tag{1}$$
$$\begin{cases} 1.0 & \text{if sentence match} \\ P(sat|o(N_{min}, N), \neg sent\ match) & \text{if } o(N_{min}, N) = 1 \\ P(sat|\neg o(N_{min}, N), \neg sent\ match) & \text{if } o(N_{min}, N) = 0 \end{cases}$$

This value can then be averaged over an arbitrary dataset, to produce the Expected Search Satisfaction Rate (ESSR). A more fine-grained model could be derived by conditioning on a combination of overlap outcomes, e.g., $P(sat|o(1, 10), \neg o(3, 5))$. However a larger data collection would be required to estimate each parameter correctly.

4.1 User satisfaction prediction on held out data

In order to validate the satisfaction model introduced in the last section, we held out 1500 voice search ratings, each labeled by 3 annotators who did not label any of the training utterances. As in Section 3, the reference rating is computed through majority voting. Table 5 compares the estimated satisfaction probability based on the model in (1) with the actual percentage of satisfactory search results. Results are computed on 1036 utterances, after removing utterances for which the reference human transcript produced unsatisfac-

Table 5: Relative user satisfaction error for sentence match and the ESSR models in Table 4, on a held out human evaluation set of 1036 utterances for which the reference transcript returned satisfactory search results. The relative error is computed as 1-predicted satisfaction/actual satisfaction.

sentence match	Predicted ESSR P(sat|o(1,10))	Predicted ESSR P(sat|o(1,3))	Predicted ESSR P(sat|o(3,5))
-21.8%	-0.9%	-1.3%	-1.1%

Table 6: Individual satisfaction variation and relative ESSR error based on ground truths derived from individual human judges, using the ESSR model in Table 4. Satisfaction variation is the difference between the mean ground truth satisfaction and the satisfaction reported by individual judges.

Human judge	Test slice	Voice searches	Satisfaction variation	ESSR P(s|o(1,10))
Judge 1	1	444	-1.8%	0.1%
Judge 2	1	470	0.4%	-2.3%
Judge 3	1	470	1.2%	-2.9%
Judge 4	2	548	2.8%	-4.7%
Judge 5	2	523	-1.6%	-0.7%
Judge 6	2	526	-1.3%	-0.6%

tory search results. Results show that the satisfaction models predict the true user satisfaction within ±1.3% relative error, with the best model predicting an ESSR 0.9% lower than the ground truth satisfaction percentage. In contrast, the percentage of exact sentence match (1-SER) is 21.8% lower than the ground truth satisfaction.

4.2 Impact of individual differences

In order to assess the impact of individual judges on our evaluation method, we evaluated the percentage of satisfaction according to a single human judge, rather than taking the majority vote. Table 6 reports the satisfaction estimated on the two slices of the test data used in Section 4.1, which was annotated by different groups of 3 judges. We find that on the same set the perceived satisfaction varies by up to 4.1% absolute in the worst case scenario. Since we remove utterances for which the reference transcript is out of domain or producing unsatisfactory results, it's important to note that this variation results from two factors: (1) whether or not an utterance is in domain and intelligible, and (2) whether the results are satisfactory.

Regarding inter-rater agreement on the full set of 3000 searches, we find that the judges agree 90.0% of the time on whether or not an utterance is in domain and intelligible (vs 50% by chance), with an average inter-rater correlation of 0.38. Regarding satisfaction judgments (1, 2 or 3), judges agrees 67.1% of the time (vs 33.3% by chance), for an average correlation of 0.65. Results are averaged over 15 judge pairs.

5. LIMITATIONS

An issue with the proposed approach is that search engines are not static over time. This can affect our evaluation method in two ways: (1) the reference search results should be collected at the same time as the test search results for the comparison to be meaningful; and (2) the satisfaction model might need to be re-estimated over time. While query/result pairs can be cached to speed up offline model evaluations, the cache needs to be updated frequently to address issue (1). Given the small number of parameters in our satisfaction model, we believe that the risk of overfitting is limited. However, significant changes in either the ASR model or the search engine could make the predicted ESSR inaccurate over time.

A second limitation is that the learned parameters are only valid for the search engine on which they were estimated. However the cost of estimating new parameters is fixed and relatively small compared to the cost of evaluating the search results produced by any ASR model update. Note that this paper focuses on product search as an example, however the methodology is content agnostic. Hence future work will assess how the method scales to other domains (e.g., text document retrieval), as well as to search engines with different levels of performance.

6. CONCLUSION

This paper has presented a voice search evaluation method which is a better predictor of user satisfaction than an exact sentence match, with no additional human annotation cost. Our model can be seen as a data-driven extension of the Google Webscore beyond the 1-best search result, which is crucial given that search engines are sensitive to the tokenization produced by the ASR. The satisfaction model predicts the Expected Search Satisfaction Rate (ESSR) within ±0.9% relative to the ground truth satisfaction on held out data, while an exact sentence match underestimates the true satisfaction by 21.8% relative. Those results suggest that ESSR is a better objective function for (a) tuning voice search models and (b) monitoring accuracy over time.

7. REFERENCES

[1] A. Franz and B. Milch. Searching the web by voice. In *Proceedings of the International Conference on Computational Linguistics*, 2002.

[2] X. He and L. Deng. Speech-centric information processing: An optimization-oriented approach. *Proceedings of the IEEE*, 101(5):1116–1135, 2013.

[3] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR*, pages 41–48, 2000.

[4] P. Jyothi, L. Johnson, C. Chelba, and B. Strope. Large-scale discriminative language model reranking for voice-search. In *Proceedings of NAACL HLT*, 2012.

[5] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, , and B. Strope. Google search by voice: A case study. In A. Neustein, editor, *Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers and Clinics*. Springer, 2010.