



Speech Transcript Evaluation for Information Retrieval

Laurens van der Werff¹, Wessel Kraaij², Franciska de Jong¹

¹Human Media Interaction, Twente University, Enschede, The Netherlands

²Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands

laurens75@gmail.com, w.kraaij@cs.ru.nl, f.m.g.dejong@ewi.utwente.nl

Abstract

Speech recognition transcripts are being used in various fields of research and practical applications, putting various demands on their accuracy. Traditionally ASR research has used intrinsic evaluation measures such as word error rate to determine transcript quality. In non-dictation-type applications such as speech retrieval, it is better to use extrinsic (or task specific) measures. Indexation and the associated processing may eliminate certain errors, whereas the search query may reveal others. In this work, we argue that the standard extrinsic speech retrieval measure average precision is unpractical for ASR evaluation. As an alternative we propose the use of ranked correlation measures on the output of the speech retrieval task, with the goal of predicting relative mean average precision. The measures we used showed a reasonably high correlation with average precision, but require much less human effort to calculate and can be more easily deployed in a variety of real-life settings.

Index Terms: evaluation, speech recognition, information retrieval, speech retrieval, rank correlation

1. Introduction

Speech Retrieval (SR) is usually implemented by using a customized information retrieval (IR) engine on the output of a large vocabulary continuous speech recognition (LVCSR) system. Traditionally, the quality of speech transcripts is measured using word error rate (WER) indicating the percentage of incorrectly transcribed words. This is an intrinsic evaluation, i.e., only the result itself is evaluated without consideration of its application. When transcripts are used for SR purposes, many errors become irrelevant due to the indexing process, which often stems all incoming terms and removes the most frequent ones. In order to anticipate these circumstances, we need to use an extrinsic, or task-based, evaluation of the transcript. The standard measure for quality of ranked retrieval systems is (mean) average precision (MAP), but it is rather laborious to determine due to the need for human-made relevance judgements. If a reference transcript is available, it is therefore desirable to have an extrinsic measure that does not require relevance judgements.

The primary task of SR is ranking speech fragments for their expected relevance towards an information need as defined in a query. Evaluation of such systems with MAP requires relevance judgements for each query/fragment pair, called *qrels*. As MAP calculation typically needs around 50 queries to stabilize and a realistic collection can span more than 10,000 fragments, it is rarely feasible to create full *qrels* for an ad-hoc collection or information need. In order to be able to optimize a speech transcript, the absolute level of retrieval performance is irrelevant; for search only the impact of transcription errors on retrieval matters. Using a reference transcript we can calculate absolute

intrinsic performance (e.g. WER) or relative extrinsic performance, allowing for the optimization of ASR in context.

In this paper we have applied existing methods of comparing ranked lists to a problem for which this has not been applied before: evaluating ASR in an SR context. For our application we define the ranking that results from a reference transcript as a ground truth, and calculate the deviation in ranking that results from transcript noise. An implicit assumption is that a perfect literal transcription is optimal from a retrieval point of view. Although it has been shown that this is not necessarily true [1], there is no reason to assume that a structural improvement can be expected from random transcript errors. The main advantages of these methods are the ability to measure the quality of the transcript for any arbitrary information need, and incorporating many aspects of the SR system, including preprocessing, use of transcript alternatives, and general retrieval strategy.

The rest of this paper is organized as follows: Section 2 provides an overview of earlier work on evaluation in the context of speech retrieval, and Section 3 explains the methods we applied in our experiments. Section 4 describes our experimental setup for our investigation into how the new method compares to WER and MAP, the results of which are given in Section 5. Finally, Sections 6 and 7 provide a conclusion and directions for future work.

2. Background

One of the earliest formal large-scale investigations into SR was done for TREC7 SDR [2]. A high correlation was found between the ranking of systems based on WER and MAP, indicating that intrinsic and extrinsic evaluation gave roughly similar results. Recognizing that this may have been serendipitous to the procedures used, several alternative ASR for SR evaluation methods were investigated. (i) (Story) Term Error Rate (TER) is vaguely similar to WER but is calculated as the sum of difference in term counts for each story divided by the total term count of the collection, thereby not requiring an explicit alignment, and counting a substitution error as an insertion plus a deletion. (ii) Stemmed and Stop Word Filtered WER is similar to WER but stop words are removed and all remaining terms are stemmed. (iii) Named Entity WER is WER calculated after all non named entities are removed from the transcriptions. Only the latter seemed to result in a slightly better correlation with MAP, whereas the other measures gave results similar to TER (i.e. quite good).

In later versions of the TREC SDR benchmarks [1] the correlation between system ranking based on WER and MAP was even higher than for TREC7, leading to the almost exclusive use of these two measures in the evaluation process. The overall level of performance even lead to the task of SR on broadcast news data being declared 'solved' [1].

In [3] the relative impact of transcript errors on SR performance by ‘error type’ was investigated. The least impact could be attributed to errors where only the count of terms was changed but not their binary presence. The complete deletion of a term from a document mostly impacted long queries, whereas the insertion of a term otherwise not present in the document was most detrimental for retrieval using short queries.

A combination of the findings in [2] and [3] could be found in the Indicator Error Rate (IER) which was introduced in [4]. It is a variation on TER where transcriptions are stopped and stemmed, and a binary error count was used. Correlation between IER and MAP was potentially higher than for WER or TER, but results were inconclusive due to the limited number of data points.

In [5] an alternative called Relevance-based Index Accuracy (RIA) was presented where the transcription accuracy was calculated directly from the resulting IR-index relevance weights. Generally this is similar to IER and SSWF-WER in that it includes stopping and stemming, but it relies on the non-linearity of the weighting function of the IR system for determining the impact of each error. In a sense this method is intrinsic as it does not include any specific information requests, but is extrinsic in relying on the indexing and pre-processing properties of the SR system that uses the transcript.

3. Correlation-based evaluation

The methods described in Section 2 used a variation on WER which included weighting or filtering of terms and errors. The implicit assumption was that the quality of an SR transcript is independent of the context it is used in. A high correlation between intrinsic and extrinsic methods only confirms the correctness of this assumption if the information need is relatively stable and well represented in the extrinsic testing.

But what constitutes a typical query depends on the type of user and the collection. For a heterogeneous collection that is searched by the general public, for example a collection of (local) radio broadcasts, anything is possible, and typically queries will be only a few words long [6]. When the collection is more specialized, for example a set of interviews with holocaust survivors and available only to professionals, queries may be much longer and less ambiguous as a consequence.

Incorporating the expected use as a variable into the evaluation of the transcript is done by looking directly at the output of an SR system, for example using MAP. Coming from the Cranfield[7] tradition, MAP uses binary relevance for query/fragment pairs. This limits the ground truth from being a ranking to a binary classification. As ASR errors not always impact relevance, MAP may not be sufficiently sensitive towards transcript noise. In addition, classifying all documents in a collection for each individual (test) query is very time-consuming, making evaluation with MAP rarely feasible for a non-benchmark task, despite some very effective strategies for reducing the required amount of judgements.

We propose using the ranking of documents that results from a retrieval task on a human-made reference transcription as a ground truth. Clearly, this is not a result that gives a maximum MAP score, but it may be equivalent to the best possible MAP score given all constraints besides the transcription noise. We can then use the correlation between ranked results lists from SR on ASR transcripts and manual transcripts as a measure for transcript quality. The main advantage of this approach over MAP is that it overcomes the need for qrels, enabling evaluation on ad-hoc queries and collections.

The most popular methods for calculating the correlation of ranked results lists are Kendall’s τ and Spearman’s ρ [8]. Both were designed for calculating ranked correlation between two lists containing the same items, but for the results of two separate speech retrieval runs this is generally not the case. We dealt with this using the solution that was proposed in [9]: all items that occurred in one list of length N , but not the other were treated as if they occurred in that list at rank $N + 1$. This potentially introduces tied ranks at position $N + 1$. For our τ -based measure we chose to resolve those using the averaging Kendall distance from [9] meaning that for a tie, the number of concordant pairs is increased by 0.5 (as opposed to 1 for a true concordant pair). For the ρ -based measure it was unproblematic.

Generally the highest ranked results are considered the most important in any IR task, whereas the tail of the results list (especially when the number of results runs in the hundreds) is often never inspected in reality. This assumption is underlying any use of MAP, which is biased towards the top of the result list. Alternatives for Kendall’s τ and Spearman’s ρ have been proposed that also put more emphasis on the top end of the lists. They are Average Precision inspired Tau (τ_{ap}) [10], see Equation 1, and Blest’s Rank Correlation Coefficient (ρ_B) [11], see Equation 2.

$$\tau_{ap} = \frac{2}{N-1} \sum_{i=2}^N \left(\frac{C_i}{i-1} \right) - 1 \quad (1)$$

$$\rho_B = \frac{2N+1}{N-1} - \frac{12}{N(N+1)^2(N-1)} \sum_{i=1}^N (N+1-i)^2 q_i \quad (2)$$

C_i represents the number of items above rank i in one list that is correctly ranked with respect to the item at rank i in the other list, q_i is the rank in the other list of the item at rank i . N is the number of results in each list.

We propose using these two ranked correlation coefficients to determine the match between the outcome of SR on a reference and an ASR transcript and expect this to closely correspond with the ‘quality’ of the transcript in the context of the specific SR system used. As we typically want to use multiple queries to assess performance, a mean value over all queries is used.

4. Experimental Setup

We used the TDT-2 English language broadcast news speech collection[12] for our experiments. Despite it being relatively old now and not particularly challenging from either an ASR or an IR point, it has a number of very desirable properties that no other current collection can offer. MAP can be calculated thanks to a full set of qrels ($\approx 4k$ relevant out of $\approx 22k$ fragments) for 100 information requests. Furthermore, the 400 hours were fully transcribed manually with 10 hours in reference quality and the remainder as closed-captions or quick transcriptions. As part of the TREC8 and TREC9 SDR benchmarks[1] several labs made a full automatic transcript for this collection, seven of these were available to us, plus one additional new transcript from 2008 (courtesy of Limsi). Three labs (Cambridge University, Sheffield University, and Limsi) submitted multiple automatic transcriptions using the same ASR system but with a different quality/speed optimization.

Several alternative story (fragment) boundaries were also added to the comparison. The boundaries were generated using the methods described in [13], with the prefix ‘F’ indicating

	TER			MAP	
	full	trec8	trec9	trec8	trec9
Limsi08	17.28	11.59	9.05	38.55	30.26
Limsi2u	19.86	13.54	10.93	37.93	29.02
CUs1u	20.18	13.68	11.10	38.24	29.00
Limsi1u	20.81	14.09	11.61	36.54	28.53
CUs1p1u	24.04	16.93	14.37	36.40	27.72
NistB1u	24.26	16.65	14.50	35.23	27.63
Shef2k	26.07	18.02	15.29	35.18	27.14
Shef1k	28.11	19.79	16.99	34.31	26.86
corr. w/MAP	0.93/1	0.86	1		

Table 1: Baseline TER and MAP performance for 8 transcriptions. TER was calculated for the whole set and only for terms that were used in the trec8/trec9 query sets. (Macro) Correlation was measured using Kendall’s τ .

boundaries that were based on fixed-length intervals, ‘WN’ the use of WordNet and ‘TT’ stands for TextTiling. Three parameter settings for each method were used, one optimal (1) and two near-optimal (2 & 3). The experiments on the automatic story boundaries used the reference transcript, the experiments on the multiple ASR transcripts used the reference boundaries.

For our experiments we custom-built an IR system using Perl and MySQL, with ranking based on the BM25/okapi ranking function[14], see Equation 3. Term Frequency for a given term/fragment pair is represented by $tf_{t,d}$, dl_d is the length of the fragment in words, dl_{avg} the average fragment length of the collection and df_t the number of fragments containing the term. Tuning parameter k_1 was set at 1.1 and b was optimized for each individual run.

$$bm25_{t,d} = \frac{tf_{t,d}(k_1 + 1)}{k_1 \times ((1 - b) + b \times \frac{dl_d}{dl_{avg}}) + tf_{t,d}} \times \log \frac{N}{df_t} \quad (3)$$

Query and document expansion were not used, as although their use typically leads to improved retrieval performance it also tends to reduce, and therefore obscure, the impact of ASR errors[1]. For the collection and queries at hand, the impact of ASR errors was already quite small, making it potentially impossible to distinguish automatic from manual transcripts for the given task.

The transcriptions were used ‘as is’ and only the 1-best transcript was considered, i.e., no confidence scores and/or transcription alternatives were included. Before indexing, all terms were stemmed using a Porter stemmer [15] and stop words were removed. No additional normalization was done. Retrieval experiments were carried out using two separate sets of queries (50 each), which we refer to as ‘trec8’ and ‘trec9’. Ranked correlations of the output of retrieval runs were calculated for the top 1000 results, the same as for calculating MAP.

5. Results

Term Error Rates ranged from 17 to 28%, see Table 1, but what those numbers do not tell is how different the transcripts were. A high overlap in errors makes TER more likely to be correlated with retrieval performance. Having the same errors but more of them, is unlikely to lead to better retrieval, whereas having different errors but the same amount could. The overlap in errors between systems from the same lab ranged from 79 to 83%, whereas the overlap with systems from other labs was 64 to 68%, regardless of overall error rate. All transcripts from the

	trec8		trec9	
	τ_{ap}	ρ_B	τ_{ap}	ρ_B
Limsi08	64.11	78.91	72.86	82.52
Limsi2u	63.98	75.69	68.05	79.01
CUs1u	64.06	75.74	68.29	79.10
Limsi1u	63.07	75.17	66.80	78.16
CUs1p1u	58.51	70.87	62.38	74.31
NistB1u	60.33	72.80	63.97	75.77
Shef2k	57.09	70.13	62.05	74.17
Shef1k	54.50	68.02	58.94	71.41
corr. w/MAP	0.93	0.93	0.86	0.86

Table 2: Ranked list correlations τ_{ap} and ρ_B between IR on reference and 8 ASR transcriptions. (Macro) Correlation with MAP was measured using Kendall’s τ .

same lab were identically ranked based on TER and MAP.

Also not shown in Table 1 is the distribution of errors, which may be important for SR purposes. All systems had an increasing proportion of errors as terms became less frequent, with 61-65% of all errors in the 50% of the transcript that was taken up by the least frequent terms, and 38-43% for the bottom 25%. The best systems had a lower proportion of errors on the most frequent terms, indicating that the higher performance benefitted frequent terms more than infrequent terms.

There was much better ASR performance for just the trec8 and trec9 query terms than for the entire set, indicating that the query terms were not representative of transcript quality (but could of course still be representative of queries of this length).

The correlation between the results of a retrieval task on the reference and the various ASR transcripts is shown in Table 2. Correlation with MAP was the same for τ_{ap} and ρ_B , equal to TER for the trec8 queries, but slightly worse for trec9. The differences between MAP of some systems was so small that it is difficult to draw any definite conclusion though: NistB1u and Shef2k were a mere 0.05 MAP apart on trec8, and Limsi2u and CUs1u were an even closer 0.02 MAP on the trec9 queries.

Automatic fragment boundaries were previously evaluated using a cost-function on this task [16], but for the nine sets of boundaries in Table 3 the correlation with MAP was non-existent ($\tau=0.16$ and 0.11). The τ_{ap} and ρ_B measures proved much more successful. Correlation in system ranking between τ_{ap} and MAP was highest for trec8, with ρ_B performing a little better on trec9 queries. Neither of these results was as good as those in Table 2. However, when looking at each story boundary *method* individually, τ_{ap} placed the various parameter settings in the correct order of merit for each of the three methods and for both query sets.

Combining the results from Table 2 and 3 gave an overall correlation with MAP for τ_{ap} of 0.96 for trec8 and 0.88 for trec9, and for ρ_B of 0.93 and 0.90.

6. Discussion and Conclusion

Evaluating transcripts in the context of speech retrieval is needed for optimization of the ASR task. WER, or the closely related TER which we used in our experiments, showed a very high correlation with retrieval performance (MAP) for the eight transcripts of TDT-2. Limitations are that WER is only properly defined for simple 1-best transcripts (no transcript alternatives and confidences), and that correlation with MAP may be reduced when calculated for a specific information need. The transcripts that were available for our experiments were all cre-

	trec8			trec9		
	MAP	τ_{ap}	ρ_B	MAP	τ_{ap}	ρ_B
F1	23.70	35.36	54.11	20.68	35.54	54.83
F2	19.54	32.79	50.14	17.83	32.70	50.63
F3	16.10	27.29	43.06	12.50	27.67	43.64
WN1	23.15	34.70	52.97	18.26	34.94	53.80
WN2	19.86	33.49	51.12	16.84	33.60	51.51
WN3	19.22	31.99	48.99	16.38	32.07	49.58
TT1	20.15	34.12	51.70	15.59	33.89	51.20
TT2	18.41	32.79	49.69	14.11	32.46	49.64
TT3	18.06	32.38	50.19	15.30	32.69	50.46
τ		0.89	0.78		0.67	0.72

Table 3: MAP, τ_{ap} , and ρ_B for results of IR with trec8 and trec9 queries, using different sets of story boundaries. (Macro) Correlation with MAP was measured using Kendall’s τ .

ated using very similar ASR systems, and their differences were mostly in the amount, and not the type of errors.

The alternatives we proposed showed a lot of promise. For the τ_{ap} and ρ_B measure, correlation for one set of queries was as high as with TER, whereas the other set scored only slightly lower. Given the Cranfield-induced limitations of MAP, the slight differences in system ranks that we saw in our results should be interpreted carefully. Both TER and the correlation-based measures can be calculated without a need for qrels, tackling one of the biggest problems with extrinsic evaluation of ASR transcripts for SR. Using correlation of results, any provisions in the SR system that attempt to reduce the impact of errors on retrieval performance are taken into account. This makes the extrinsic methods potentially much more useful for optimization in the context of SR.

One of the main attractions of our correlation-based methods is the reduction of workflow in comparison to a design that requires relevance judgements as for MAP. In this work, we have used a full (although ‘sloppy’) transcription of the TDT-2 collection. But creating a full manual transcription of a 400 hour collection is not particularly appealing, and defies the purpose of the exercise which is to optimize the creation of an automatic transcription. More research needs to be done to determine what amount of manual transcripts is needed for these measures to stabilize.

The correlation-based methods were much more successful at ranking fragment boundary sets for SR performance than the previously used cost function. The most attractive method overall was τ_{ap} . Its ranking of systems was highly correlated with that of MAP for various ASR transcripts, it was perfectly correlated with regard to parameter settings for automatic boundary generation, and was reasonably good at ranking boundary generating methods. In conclusion we expect the methods we propose to apply to the evaluation of ASR transcripts in an SR context are a useful addition to existing measures such as TER and may be preferable in a number of scenarios.

7. Future Work

There is clear added value for the proposed transcript evaluation methods when compared to TER with regards to transcript alternatives, e.g., in the form of lattices, and confidence scores. More experimentation is needed to determine which SR settings would benefit most from adoption of the proposed measures.

Once better understood, they could be used to improve ASR system settings, for example by tuning language and/or acoustic

models in view of optimizing retrieval performance rather than transcript quality. This could turn out highly beneficial for collections for which qrel generation and therefore evaluation with MAP is unfeasible, such as collections from the cultural heritage domain, and other collections that need more tuning than broadcast news data.

8. Acknowledgements

The research reported on here was funded by the research project CHoral¹, part of the NWO-CATCH² program. We would like to thank the LIMSI for kindly providing us with a full transcription of the TDT-2 speech corpus.

9. References

- [1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, “The TREC spoken document retrieval task: A success story,” in *Proc. of RIAO: Content Based Multimedia Information Access Conference*, Paris, France, 2000.
- [2] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, V. M. Stanford, and B. A. Lund, “TREC-7 spoken document retrieval track overview and results,” in *Proc. 7th Text Retrieval Conference TREC-7*, 1998.
- [3] A. Singhal and F. Pereira, “Document expansion for speech retrieval,” in *Proc. SIGIR ’99*. ACM Press, 1999, pp. 34–41.
- [4] W. Macherey, J. Viechtbauer, and H. Ney, “Probabilistic aspects in spoken document retrieval,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 115–127, Feb. 2003.
- [5] L. B. van der Werff and W. F. L. Heeren, “Evaluating ASR output for information retrieval,” in *Proc. of the ACM SIGIR SSSC Workshop*, 2007, pp. 7–14.
- [6] M. Kamvar and S. Baluja, “A large scale study of wireless search behavior: Google mobile search,” in *Proc. of the SIGCHI conf. on Human Factors in comp. syst.* ACM, 2006, pp. 701–709.
- [7] E. Voorhees, “The philosophy of information retrieval evaluation,” in *Evaluation of Cross-Language Information Retrieval Systems*. Springer Berlin / Heidelberg, 2002, pp. 143–170.
- [8] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*. Oxford University Press, 1990.
- [9] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing top k lists,” in *Proc. of the 14th ACM-SIAM symp. on Discrete algorithms*. Philadelphia, PA, USA: Soc. for Ind. and Appl. Math., 2003, pp. 28–36.
- [10] E. Yilmaz, J. A. Aslam, and S. Robertson, “A new rank correlation coefficient for information retrieval,” in *Proc. of the 31st ACM SIGIR conf.* New York, NY, USA: ACM, 2008, pp. 587–594.
- [11] D. Blest, “Rank correlation: an alternative measure,” *Australian and New Zealand Journal of Statistics*, vol. 42, pp. 101–111, March 2000.
- [12] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel, “The TDT-2 text and speech corpus,” in *Proc. of the DARPA Broadcast News Workshop*, 1999, pp. 57–60.
- [13] L. B. van der Werff, “Story segmentation for speech transcripts in sparse data conditions,” in *Proc. of the 2010 ACM multimedia SSSC workshop*. ACM, 2010.
- [14] K. S. Jones, S. Walker, and S. E. Robertson, “A probabilistic model of information retrieval: development and comparative experiments,” in *Information Processing and Management*, 2000, pp. 779–840.
- [15] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 14, no. 3, pp. 130–137, 1980.
- [16] G. Doddington, “The topic detection and tracking phase 2 (TDT2) evaluation plan,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

¹<http://hmi.ewi.utwente.nl/choral>

²<http://www.nwo.nl/catch>