Evaluating ASR Output for Information Retrieval

Laurens van der Werff
University of Twente
PO Box 217, NL-7500AE
Enschede, The Netherlands
laurensw@ewi.utwente.nl

Willemijn Heeren
University of Twente
PO Box 217, NL-7500AE
Enschede, The Netherlands
w.f.l.heeren@ewi.utwente.nl

ABSTRACT

Within the context of international benchmarks and collection specific projects, much work on spoken document retrieval has been done in recent years. In 2000 the issue of automatic speech recognition for spoken document retrieval was declared 'solved' for the broadcast news domain. Many collections however, are not in this domain and automatic speech recognition for these collections may contain specific new challenges. This requires a method to evaluate automatic speech recognition optimization schemes for these application areas. Traditional measures such as word error rate and story word error rate are not ideal for this. In this paper, three new evaluation metrics are proposed. Their behaviour is investigated on a cultural heritage collection and performance is compared to traditional measurements on TREC broadcast news data.

General Terms

Automatic Speech Recognition, Spoken Document Retrieval, Lattices, Evaluation

1. INTRODUCTION

Several developments in recent years have led to an increased interest in improving access to spoken word collections. The reduced cost and increased capacity of random access media (e.g., harddrives), combined with the increased speed of Internet connections, means that it is now quite feasible to access such collections online. In contrast to these technological opportunities stands the reality of current practice: many existing collections have not been properly digitized yet since this requires a lot of manual effort. Those that have been digitized are often not searchable for a variety of reasons, ranging from intellectual property issues to technical and implementation issues.

Searching in spoken content implies the application of information retrieval (IR) techniques to speech. Since searching in speech directly is unfeasible, a more computer-processable representation has to be used. From an (automatic) indexing perspective, spoken word collections can be approached in several ways based on the amount of available collateral data. Collections that are up to a few hundred hours in size can usually be made accessible through some human effort: either by labelling segments of speech with keywords and named entities or by manually creating a full transcription. This can then be automatically aligned to the audio using standard Viterbi techniques [20] and indexed as any other textual document. When an audio collection is too large to be disclosed manually, it must be done using a more or less automated process. In such cases it is expected that an automatic speech recognition (ASR) system can be used to provide a full, though imperfect, transcription of the audio.

Of great importance for the accessibility of a spoken document collection is the quality of the index. The quality of an index based on ASR output will be highly dependent on the characteristics of the speech. Since ASR is probabilistic and based on models that are estimated from statistics, performance of ASR is determined largely by the match between those models and the speech that is processed. Spontaneity and noise typically cause problems for ASR systems due to the fact that they make the speech signal less predictable and so by definition reduce the match. ASR therefore tends to perform best on material that is generated under highly controlled circumstances, for example broadcast news (BN) or dictation. Many of the collections that are considered interesting are not of this type, such as historical audio or oral history collections. These may contain noisy spontaneous speech or highly accented speech by non-professional speakers, often recorded under suboptimal conditions using old-fashioned equipment. These circumstances typically cause a doubling of the number of ASR errors and thus reduce the reliability of the automatically generated transcription.

Many optimization methods for ASR on noisy and/or spontaneous speech have been extensively studied in the past [5]. Most of these studies have employed ASR as a 'dictation machine', meaning that the primary task of the system was to generate a literal transcription of every word that was uttered. Traditionally, the performance of such ASR systems is measured using the word error rate (WER). In the context of spoken document retrieval (SDR), ASR is not so much a dictation machine as it is a means to generate some representation that is suitable for building an index. The literal transcription is just a (potential) by-product of this process. WER is a flawed optimization criterion for ASR in this context because (i) it is only defined as such on a (literal) transcription and can therefore not be calculated on ASR output such as n-best lists or lattices, and (ii) IR performance depends not only on the *amount* of errors but also on the *type* of errors.

Performance of IR systems is typically measured using the mean average precision (MAP), a score that is calculated based on the amount of relevant documents found for some set of queries, the amount of non-relevant documents that is produced and their ranking. Calculating such a score can only be done using an evaluation platform that contains ground-truth (i.e. human) relevance judgments for a set of queries and documents. When applying ASR to a collection for which such a platform exists, the MAP should be used as an optimization criterion.

In practice, IR evaluation platforms are only readily available for a limited amount of collections. When optimizing the ASR component of an SDR system for a collection for which no matching evaluation platform can be found, developing a new evaluation set requires a prohibitive amount of work. Instead, some ASR for IR optimization criterion is needed that can be used to predict the MAP, or at least the relative improvement in MAP, for collections where this score cannot be calculated. In this work, three new performance measures for ASR in an SDR context are introduced.

This paper is organised as follows: Section 2 touches on some previous efforts to find the relationship between ASR and IR performance, strengthening the argument that WER is not a good criterion for optimizing ASR in an IR environment. Section 3 first explains the workings of an SDR system and why current evaluation metrics for ASR can be problematic in this context. Then three new performance measures will be proposed that are more appropriate versions of the traditional measures WER, Story WER (SWER) and Out-Of-Vocabulary (OOV) rate. It is argued that the ASR output can only be assessed properly when some particular characteristics of the IR system are incorporated into the evaluation. The behaviour of the measures in combination with standard IR techniques is investigated in Section 4 and in Section 5 a comparison is made between the traditional ASR measures and the new ones on a TREC BN collection. Finally, Section 6 contains some conclusions and gives suggestions for future work.

2. RELATED WORK

The performance of ASR in the context of IR has been studied for many years, mainly in the context of TREC since 1997 [21]. In [4] it was noted that there is a high correlation between WER (actually SWER) and retrieval performance as measured with MAP. This correlation was even higher when instead of SWER a Named Entity SWER (NE-SWER) was used, measuring exclusively the named entity performance of the ASR system. In [9] some experiments were done with Term Error Rate (TER) as a performance measure. A high correlation was found between the TER and the MAP score of the systems, however no such clear relationship was found with the R-precision score. Since SDR performance on ASR based transcriptions was only marginally worse than on human transcriptions, ASR-based indexing was considered ‘solved’ for the BN domain [3].

In [19] the IR performance of indexes based on different types of ASR output was evaluated. The incorporation of the sentence structure through the use of n-best lists was found to be superior to using individual word probabilities from lattices. Using 1-best output was found to be inferior to using either n-best or lattice representations. The IR weights were calculated by combining relevance and ASR confidence into a single probabilistic measure. The effect of the choice of ASR output type on overall retrieval performance was measured by running and evaluating a predefined set of queries on the resulting index and comparing MAP score.

More recently, research has been done on optimized indexing from ASR lattices for improved IR performance, for example through multi-word queries [1] or through combination of multiple lattice hypotheses [14]. Both techniques gave rise to some improvement.

This previous work suggests (i) that IR performance is dependent on ASR performance, (ii) that indexing from lattices or n-best lists can improve IR performance and (iii) that the way that these enriched outputs are exploited needs to be optimized.

3. EVALUATING SDR

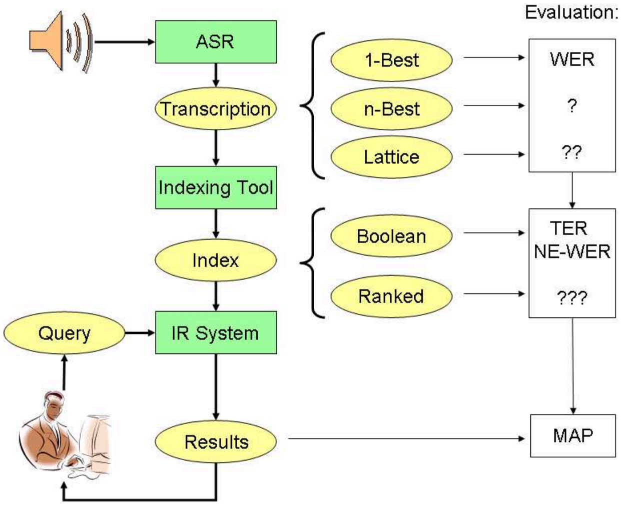

Figure 1: Anatomy of an SDR system

3.1 Anatomy of an SDR system.

A typical SDR system will contain at least the following three main components: an ASR engine, an indexing tool and an IR system. Figure 1 gives an overview of such an SDR system. The ASR engine takes as its input the audio containing the speech and produces a transcription. This can then be processed into some index representation by the indexing tool. Finally, the user enters queries into the IR system which will produce the relevant audio fragments based on the index.

These three components should work together in such a way that the final retrieval results are optimal given the user's query. MAP is the standard method for evaluating this, so optimizing the individual components for the best MAP score is the most efficient way of improving system performance. Calculating the MAP score however is not always possible, it requires some evaluation platform which, as was mentioned in Section 1, is very time consuming to produce.

Although the inability to calculate a MAP score for most collections might limit optimization and customization of an IR component, it may still be possible to optimize the ASR system and the indexing tool. As mentioned in Section 2, there is a fairly strong correlation between ASR performance as measured with WER and the MAP score. However, as will become clear in Section 5, optimisation of the ASR component for minimal WER does not automatically lead to an SDR system with a higher MAP. Proper evaluation of the ASR output and/or the results of the indexing tool is therefore crucial for optimization of SDR systems for collections that do not allow for the calculation of MAP.

3.1.1 Transcription Types

The output of the ASR engine can take several forms. Traditionally, in dictation type applications, the 1-best output is used. It represents the sequence of words that, based on the acoustic and language models used as well as pruning parameters, gives the highest likelihood for a fragment of speech. A 1-best output normally does not contain any scoring information for individual words, meaning that confidence in its correctness is equal for each word in the transcription. Evaluation of the 1-best output is done using WER, calculated using the following equation:

$$WER = \frac{S + I + D}{N}$$

Where S, I and D represent the number of substitutions, insertions and deletions as determined through a dynamic programming, minimum Levenshtein distance function (weights: 4, 3 and 3)[13] alignment of reference and hypothesis transcription. N is the total number of words in the reference.

Alternatively, an ASR engine can produce an n-best list or a lattice structure as its output. Both of these types of output contain multiple transcriptions for the audio and may also contain some form of confidence scoring. The main difference between them is that n-best lists contain only full transcription alternatives, i.e. full sentences, while lattices contain alternatives on a word-by-word basis. A lattice structure is a relatively compact representation of the search space of the ASR engine and can be expanded into an n-best list. Lattice output is typically used as an intermediate representation that is then postprocessed/rescored into a 1-best output which in turn can be evaluated using WER. When lattice or n-best output has to be evaluated directly, no useful metrics are available.

3.1.2 Indexing

The index of an IR system links words and/or concepts to specific documents (or speech fragments in the case of SDR). In IR that is based on textual documents, the underlying data on which the index was made is, in principle, reliable. When the index is based on ASR output, the reliability of the index may suffer as a result of transcription errors. Since final retrieval performance is directly dependent on the index and only indirectly on ASR performance, evaluation of ASR output by measuring the impact of the errors on the index should, at least in theory, be more indicative of IR performance than evaluation of the ASR output by itself. Evaluation of an ASR-based index can be done by building an index both on a reference transcription and on the ASR output and comparing the two.

For a Boolean retrieval system, each index term represents an unambiguous set of documents: those that contain it. Measuring the impact of ASR errors on the index is therefore a matter of counting these errors, for example using the term error rate (TER) as proposed in [8]:

$$TER = \frac{\sum_w |A(w) - B(w)|}{W}$$

Where W is the total number of words in the reference and $A(w)$ and $B(w)$ represent the number of times word w occurs in the reference A and the transcription B, thereby modeling a traditional substitution as two errors. Since the number of occurrences of a word is of no importance in a Boolean system – a document is either a member of a set or it is not – a unique term error rate (UTER) value may be more appropriate. This can be calculated by using $A(w)$ and $B(w)$ only to indicate the presence (value=1) or absence (value=0) of word w in the document.

The family of ranked retrieval models is characterized by the inclusion of a – usually statistically motivated – weighting scheme on the index terms. Such a scheme is typically based on some form of term frequency (tf) and document frequency (df) combination. Several approaches exist for exploiting and calculating these measures, for example the Vector Space Model (VSM) [17] and Okapi [11].

Measuring the impact of ASR errors in a ranked retrieval environment is not simply a matter of counting, since errors now impact weights in a complex manner. A deletion of a term will decrease its tf for that document, but will also decrease the df that is calculated over the whole set, thereby increasing the weight for this term in all other documents.

The TER can be adapted as proposed in [7], so that the error count of each term is multiplied by an individual weight. This can be used to simulate the non-uniform impact of ASR errors, but finding a suitable weighting function may be quite difficult and the total error is still determined by simply counting the number of insertions and deletions.

3.2 SDR Evaluation Metrics

In the systems that were enrolled in the TREC benchmarks, ASR performance was measured using the WER (all systems used 1-best ASR output only)[4]. By comparing the ranked retrieval IR performance of the systems on each of the various ASR outputs, a correlation between ASR performance and retrieval performance could be established. As it turned out, the correlation coefficient in the TREC-7 systems between WER and MAP was 0.87, meaning a significant correlation. The NE-SWER showed an even higher correlation with the MAP at 0.91. Although this might validate the conclusion that WER is a good measure for predicting relative IR performance, there is more to this.

The ASR components of all systems that took part in this evaluation were optimized for the same evaluation metric:

WER, a measure that does not differentiate between errors on content words or on stopwords. Since all ASR systems had the same basic layout, it could be argued that the performance of these systems will not differ very much in a qualitative way, so pure quantitative analysis could be sufficient. Comparing WER with NE-SWER, the *relative* performance of all of the systems stayed the same, except for one. This was precisely the system that had shown lower relative MAP scores than would have been predicted from its WER, but it showed an NE-SWER that was in line with its MAP score. This increased the overall correlation coefficient and supports the notion that a qualitative measure may be useful for ASR evaluation in an IR context.

Quantitative analysis of ASR performance is only indicative of retrieval performance if this was also the criterion used for optimizing the ASR system, as is the case in most dictation type applications. When optimizing an ASR system for a different application, leading for example to an increased performance on named entities at the cost of performance on stopwords, the WER may no longer be a good indication of relative retrieval performance. When an index is built using n-best or lattice output, the WER cannot even be calculated as such. This is further reason to conclude that different ASR performance metrics are required for SDR. The following paragraphs will introduce three such measures.

3.2.1 Boolean Index Accuracy

In a Boolean retrieval system, the index *is* the system, since queries are simply a way of selecting documents from a combination of sets that are entirely defined by the index. In the context of such a system, measuring the quality of the index is a matter of calculating the TER and is therefore quite straightforward.

When an index is created based on n-best or lattice ASR output, the number of terms that are associated with a document becomes quite variable. When only words for which confidence in the ASR correctness is very high are included, this leads to a relatively small number of associated terms, while inclusion of several alternatives for some sentence-positions will increase the amount of terms.

In practice, due to the possibility of creating relatively complex indexes from lattices or n-best lists, the TER (or UTER) value may become much larger than 1 (or 100%), making it difficult to interpret unambiguously. For example, is an 'empty' index with a TER of 1 better than a relatively large index with a TER value of 1.5? It would be preferable to always indicate the performance with a number between 0 and 1, where 0 would mean no match between hypothesis and transcription, while 1 would indicate that the hypothesized index is equal to the reference index. The Boolean Index Accuracy (BIA) is such a measure:

$$BIA = \left(1 - \frac{D}{N_{ref}}\right) * \left(1 - \frac{I}{N_{index}}\right) \tag{1}$$

Where D is the number deletions, meaning terms that are in the reference, but not in the hypothesis, while I is the number insertions, meaning terms that are in the hypothesis but not in the reference. N_{ref} is the number of terms in the reference, while N_{index} contains the number of terms in the index. Terms are considered unique for a particular story (or retrieval unit). Equation 1 is made up of two parts: the first bracketed part indicates the coverage of the index, i.e. the fraction of the words in the reference transcription that can be found in the index. The second bracketed part indicates its correctness, i.e. the fraction of the words in the index that is also found in the reference transcription.

3.2.2 Ranked Index Accuracy

In a system of ranked retrieval, the index contains weights for each indexable term in each document. These weights determine the ranking and therefore define the system. Measuring the similarity in weights between the hypothesized index and the reference index can be done using the standard VSM [17]. In this model, the index can be represented as a vector, with the indexed terms as vector dimensions and the weighting scores as vector lengths. By calculating the vector inner product of the normalized vectors, the similarity of two indexes can be determined. This property can be expressed in the RIA measure that is calculated as follows:

$$RIA = \frac{\sum_{k=1}^{m} d_k \cdot q_k}{\sqrt{\sum_{k=1}^{m}(d_k)^2} \cdot \sqrt{\sum_{k=1}^{m}(q_k)^2}} \tag{2}$$

Where m is the combined number of terms in the indexes and d_k and q_k represent the weight of term k in the reference index d and the hypothesis index q. The Ranked Index Accuracy (RIA) represents the similarity between two indexes on a scale of 0 to 1, with 1 meaning that the indexes are identical.

3.2.3 ROOV

In optimizing ASR, the lexicon and language model are of vital importance. It is therefore useful to measure OOV rate, i.e. the percentage of words in the audio that is not included in the lexicon of the speech recognition system. In principle, the number of OOV terms is independent of the ASR output and also independent of the type of index that is made. However, since in SDR the ASR system is no longer a dictation machine and not all terms are treated equally, this measure should be adapted somewhat.
Traditionally, the OOV rate is calculated by dividing the number of OOV terms by the total number of terms in the reference.

$$OOV = \frac{\#OOV\ terms}{\#terms} * 100\%$$

Within a Boolean retrieval environment, one only needs to divide the number of unique OOV occurrences by the number of unique indexable terms to calculate the unique OOV (UOOV):

$$UOOV = \frac{\#unique\ OOV\ terms}{\#unique\ indexable\ terms} * 100\%$$

Within a ranked retrieval environment, the OOV can be calculated by dividing the total mass of all weights of the OOV terms by the sum of all the weights of the (reference) index, resulting in the retrieval OOV (ROOV):

$$ROOV = \frac{\sum Weight_{OOV\ terms}}{\sum Weight_{index}} * 100\% \tag{3}$$

When optimizing the ASR lexicon for minimal OOV, the best strategy is to include only the most frequent words in the lexicon, either estimated on a subset of the collection or on an external text corpus. When the lexicon is being optimized for ROOV, a different strategy must be chosen: for example including those words that have the highest expected weights.

3.2.4 Example of evaluation measures

When lattice or n-best ASR output is used for generating an index, the size or complexity of the index is variable: it is possible to include more or less terms from the lattice or n-best list in the generation of the index[2]. Inclusion can be done on the basis of many criteria, and weights may be adjusted accordingly. Figure 2 shows the values of the various performance measures for indexes where a variable number of terms from a lattice ASR output are included. The inclusion criterion in this case was the posterior probability.

The collection used for generating this graph contains radio recordings with (Dutch) noisy spontaneous speech, hence the relatively poor absolute performance when compared to typical results on BN type data (as found in Table 2). The total duration of the audio was approximately 220 minutes, divided into 34 stories containing an average of 1154 words per story. ASR was performed in a single pass, using BN optimized acoustic and language models and a lexicon of 65k words. This led to a WER of around 55%. For the RIA results, a $tf * log(idf)$ score was used for calculating weights. The BIA and RIA scores were calculated with an index based on a human-made transcription of the audio as the reference.

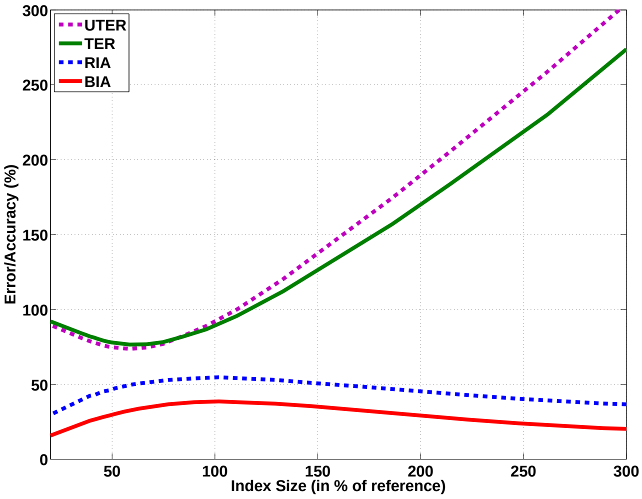

Figure 2: Comparison of performance measures for various index sizes. The index size is shown relative to the size of the reference index.

When an index is created for a certain collection, it makes sense to select the size that results in the highest similarity to the reference index. The performance of the system can thus be characterized in a quantitative manner by the maximum value of the BIA curve and in a qualitative manner by the maximum value of the RIA curve, in this case 0.39 and 0.55 respectively.

4. IR STRATEGIES AND THE ASR BASED INDEX

The evaluation measures introduced in Section 3 can be used to generate a performance number for both quantitative(BIA) as well as qualitative(RIA) evaluation. In order to show how these numbers are affected by more or less standard IR techniques, some experiments were performed with stopword filtering and stemming. The same data as in Section 3.2.4 was used.

4.1 Stopwords

In IR applications it is standard practice to filter stopwords from the index. These are words that are very common, have little or no meaning by themselves and will therefore not help in identifying relevant documents. Stopping of the most frequent words leads to a reduction in index size of up to 50% without impacting retrieval performance [18]. There is no real consensus as to what is the best size for the stopword list, but for Dutch, lists in the range of 50 to 1500 words can be found.

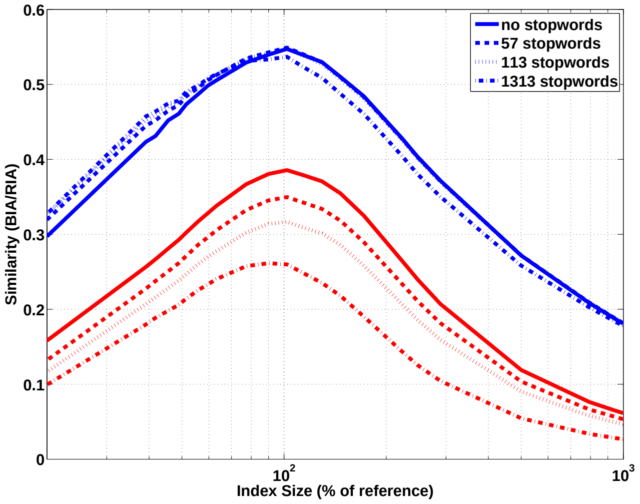

Figure 3: Index similarity after application of various stopword lists. Blue (upper) curves show RIA scores, red (lower) curves show BIA scores.

Figure 3 shows the performance curves of an ASR lattice-based index on a collection where stoplists of various sizes have been applied. The horizontal axis (index size) has been plotted on a logarithmic scale for clarity reasons. The graph clearly shows how the BIA value is impacted by using a stopword list, indicating that in this collection, ASR performance on stopwords is different (better) than on content words. The RIA measure is relatively stable, confirming that the stopwords have low relevance and ASR errors on these words may therefore have a limited impact on retrieval performance. Optimal index size seems to be at around 100% of the size of the reference index. This was to be expected, since the criterion for inclusion of terms in the index – the posterior probability – was the same criterion that was used by the ASR engine for selecting the 1-best path. The ASR engine was setup for minimal WER, which generally occurs

Stop- words	Indexable terms	U-Indexable terms	max. BIA	max. RIA	OOV (%)	UOOV (%)	ROOV (%)
0	39237	13442	0.39	0.55	4.0	7.3	11.3
57	23190 (-41%)	11980 (-11%)	0.35	0.55	5.9	7.9	11.5
113	17969 (-54%)	10811 (-20%)	0.32	0.55	7.6	8.7	11.8
1313	10488 (-73%)	6819 (-49%)	0.26	0.54	13.0	13.8	15.5

Table 1: Stopword statistics and index quality.

when the 1-best transcription is roughly the same length as the reference transcription.

Table 1 shows some statistics for this collection before and after applying the stopword lists. Although the total number of indexable terms in the transcriptions can easily be reduced by more than 50%, the number of unique indexable terms reduces much more slowly, so the reduction in index size will be less dramatic. The table also shows the various OOV measures as described in Section 3.2.3. The traditional OOV value of 4%, though not low, seems acceptable. However, when a stopword list is applied, it becomes clear that OOV rate of potential query terms in this particular SDR system is relatively high. ROOV seems to be the most robust measure, indicating more or less how much 'information' from the audio cannot be retrieved due to OOV issues. More on OOV rates and specific issues for Dutch can be found in [15].

4.2 Stemming

Both [6] and [12] found that using a Porter Stemmer [16] for Dutch did not significantly improve IR performance, but did not reduce performance either. [12] showed that a performance increase could be obtained by using more advanced algorithms, including compound splitting. It is not the aim of these experiments to build an optimal stemmer/splitter for Dutch, but merely to investigate the impact of such techniques on the quality of an index derived from an ASR run. The impact on the quality of the index as measured with BIA and RIA, using an implementation of the Porter Stemmer for Dutch is evaluated here.

Figure 4 shows the performance curves for an index based on the same ASR lattices, with and without stemming applied. Although previous studies indicated that the Porter stemmer may not improve IR performance for Dutch textual documents, these results show an increased similarity between the ASR based index and the reference index. Applying the stemmer increased RIA by 3.3% and BIA by 9.7% relative. It would therefore be interesting to further investigate whether the Porter stemmer can be beneficial for Dutch SDR, even though it is not for traditional Dutch IR.

5. COMPARISON TO OTHER MEASURES

To investigate whether the RIA and BIA measures are indeed useful for predicting retrieval performance in an SDR system, a complete IR evaluation platform must be used. Evaluations should be done based on several distinct ASR outputs. For the WER measure, this has been done for the TREC9 SDR track [3] by both Cambridge University (CU) and the University of Sheffield for seven different ASR runs. The results can be found in Table 2. Their retrieval results as well as the ASR outputs are publicly available from NIST and could therefore be used for comparisons with the values of RIA and BIA.

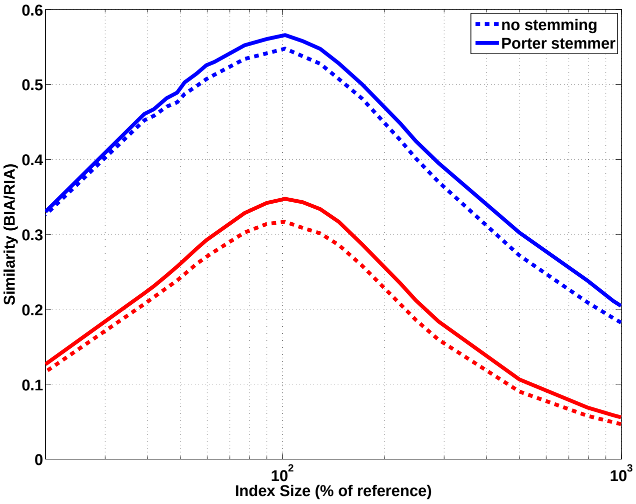

Figure 4: Performance with and without stemming applied. Blue (upper) curves show RIA scores, red (lower) curves show BIA scores.

RIA and BIA scores were calculated after applying a stopword list and a Porter stemmer to the data, as was done by both CU and Sheffield systems. RIA scores were based on weights that were calculated using the method described in [10] with constants set to the values that were reported by those labs. Table 2 shows only the RIA values based on the CU weight calculation settings, the Sheffield RIA values (not shown) were very similar. Some complications arose, possibly leading to a suboptimal calculation:

- Reference transcriptions with >10% WER were used to calculate RIA, whereas WER was estimated on a 10h subset of checked references with 0% WER
- No lab-specific normalization scripts were available, only the supplied TRANFILT tool could be applied
- Postprocessing techniques could only be approximated from the system descriptions; actual stemmers and stopword lists were not available
- Indexes were generated assuming the Story Known condition, while recognizer results and MAP scores for cross-site evaluations were only available for the Story Unknown condition

transcription	WER	SWER	RIA	BIA	CU	Sheffield
human ref	10.3	11.0			0.4402	0.4180
cuhtks1p1u	27.6	25.1	0.695	0.500	0.4044	0.3576
cuhtks1u	22.0	19.6	0.732	0.549	0.4299	0.3727
limsi1u	22.8	19.7	0.726	0.540	0.4019	0.3862
limsi2u	22.3	18.8	0.736	0.546	0.4162	0.3968
nist2000b1u	27.3	23.6	0.699	0.505	0.4075	0.3837
shef1u	33.1	28.3	0.674	0.452	0.3958	0.3919
shef2u	30.4	25.6	0.693	0.478	0.3983	0.3931

Table 2: TREC cross system results; RIA scores are based on CU parameters, the final two columns show MAP scores.

	CU	Sheffield
WER	-0.760	0.133
SWER	-0.721	-0.043
RIA	0.759	0.036
BIA	0.769	-0.132

Table 3: Correlation coefficients for MAP vs. ASR performance metrics.

Sheffield performed worse than CU on all transcription sets, however, they seemed to perform relatively well on their own transcriptions as compared to those from other sites. The best transcription set as determined by WER (cuhtks1u) led to the second worst retrieval result for this group, while the best ASR set according to SWER (limsi2u) gave the best retrieval performance. In general, no correlation between any of the measures used here and the Sheffield scores was found, nor does there seem to be any obvious correlation between the Sheffield scores and the CU scores (see Table 2).

The CU system performance showed a significant correlation with ASR quality (see Table 3). Still, the differences in retrieval performance were quite small, indicating that much of the reductions in SWER are negated by retrieval techniques such as query expansion. When Story ACCuracy (SACC) is defined as 100-SWER, its relative improvement between the best and worst transcriptions is 13.2%, the improvement in RIA is 9.2% but final retrieval performance only improves by 5.2%. RIA therefore seems to be a better predictor of retrieval performance than SACC.

All the ASR error measures used here are highly correlated (not shown). BIA is highly correlated with WER, because the index size is always within 10% of the reference. Correlation of RIA and BIA with retrieval performance is similar to their traditional error measure counterparts. The items mentioned earlier prevented the calculation of more precise RIA values, something that should not be a problem if the actual indexing software were available, as would be the case when developing ones own SDR system.

Although the CU retrieval results showed a significant correlation with WER, the Sheffield results did not. A possible cause for this lack of correlation for the Sheffield system might be that their IR component was specifically optimized for use on their own output, for example through tuning of the query expansion to the ASR lexicon or through certain post-processing techniques.

To neutralise for the effects of a better match between IR and ASR through circumstances that could not be reproduced in our calculation of RIA, comparisons were made between two different ASR runs that were produced by the same site. Three sites submitted an alternative ASR run: Cambridge, Sheffield and Limsi. When comparing retrieval performance on two ASR runs that were generated within the same site, the 'best' transcription scored consistently higher in both IR systems. Table 4 shows the performance difference in the CU and Sheffield systems between two transcriptions from the same lab.

The CUHTKS1P1U transcription from CU had an accurracy that is 7.2% lower than their CUHTKS1U version. The RIA value was 5% lower while the MAP reduced by 5.9%. The difference in RIA value for the Sheffield system in this case was 5.1%, slightly higher than for the CU system, and the MAP reduced by 4.1%. A similar trend can be found when comparing ΔACC and ΔRIA in Table 4 for the other lab's transcriptions. ΔRIA turned out to always be a better predictor of ΔMAP score than ΔACC. When the ΔRIA is compared to $\Delta SACC$ the difference was smaller, but overall still favored RIA as a predictor for MAP.

Finally, RIA10h was calculated on a ten hour subset of the reference transcription that was manually corrected (the same subset that was used to calculate the ACC numbers). It proved to be a slightly better predictor than RIA in this comparison for MAP of the CU system, but slightly worse for the Sheffield system. This indicates that RIA can also be used if a reference transcription is available for only a relatively small part of the collection.

If more details had been available of the systems that were used in this comparison, a better estimation of the RIA score could have been made, possibly leading to a higher correlation between RIA and MAP scores.

6. CONCLUSION & FUTURE WORK

In this paper, the issue of how to evaluate ASR output for use in SDR systems was raised. To avoid the use of a prohibitively expensive full IR evaluation platform, the suggestion was made to evaluate just the ASR-derived index by comparing it against an index made on a reference transcription. Three evaluation measures were introduced: (i) BIA for evaluating the errors in a purely quantitave manner, (ii) RIA for a weighted evaluation and (iii) ROOV for a weighted measure of OOV rate.

			CU			Sheffield		
Site	ΔACC	$\Delta SACC$	ΔRIA	$\Delta RIA10h$	ΔMAP	ΔRIA	$\Delta RIA10h$	ΔMAP
Cambridge	7.2	6.8	5.0	**5.5**	5.9	**5.1**	5.6	4.1
Limsi	0.6	1.1	1.4	**1.6**	3.4	**1.5**	1.0	2.7
Sheffield	3.9	3.6	**2.7**	**2.7**	0.6	**2.7**	**2.7**	0.3

Table 4: Predicted and actual performance difference of CU/Sheffield system between two ASR transcriptions from the same site; RIA is calculated on the reference transcription, while RIA10h is calculated on the manually checked 10h subset. All values are percentages.

These measures were applied to a test set with noisy spontaneous Dutch speech. Results were encouraging and in line with expectations both for performance with stopword lists as well as for performance with stemming. When a comparison was made between RIA and the more traditional WER on a set of BN data from the TREC SDR benchmarks, the new measure was significantly better at predicting changes in retrieval performance, despite the fact that its calculation was hampered by a lack of details about the IR system used.

RIA, BIA and ROOV scores can be calculated on a subset of an audio collection as is usually also done for WER estimation. The most important limitations are that the test collection must be large enough for accurate weight estimation and that the audio included in the test collection is representative for the ASR performance of the full set.

As future work, to better estimate the correlation between ΔRIA and ΔMAP, a comparison should be made in the context of an SDR system that includes a full evaluation platform. The various measures can then be compared at more stages of ASR optimization than was the case with the BN data from TREC. It would also be interesting to see if it is possible to somehow include the effects of query expansion into the measure.

7. ACKNOWLEDGMENTS

The research reported on here was funded by the research project CHoral[1], part of the NWO-CATCH[2] program.

8. REFERENCES

[1] C. Chelba and A. Acero. Position specific posterior lattices for indexing speech. In *Proceedings of ACL*. Microsoft Research, 2005.

[2] C. Chelba, J. Silva, and A. Acero. Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21(3):458–478, 2007.

[3] J. Garofolo, J. Lard, and E. Voorhees. Spoken document retrieval track slides. 2000.

[4] J. S. Garofolo, C. G. P. Auzanneic, and E. M. Voorhees. The trec spoken document retrieval task: A success story. In *Proceedings of RIAO*, 2000.

[5] Y. Gong. Speech recognition in noisy environments: a survey. *Speech Communication*, 16(3):261–291, 1995.

[6] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for european languages. *Information Retrieval*, (7):33–52, 2004.

[7] S. Johnson, P. Jourlin, K. S. Jones, and P. Woodland. Spoken document retrieval for trec-8 at cambridge university. In *NIST Special Publication 500-246*, pages 197–206, 2000.

[8] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. The cambridge university spoken document retrieval system. In *Proceedings of ICASSP '99*, pages 49–52, 1999.

[9] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. Spoken document retrieval for trec-7 at cambridge university. In *Proceedings of TREC-7*, pages 191–200, 1999.

[10] S. E. Johnson, P. Jourlin, K. S. Jones, and P. C. Woodland. Spoken document retrieval for trec-9 at cambridge university. *NIST Special Publication 500-249*, pages 117–126, 2000.

[11] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. parts 1 and 2. *Information Processing and Management*, 36(6):779–840, 2000.

[12] W. Kraaij and R. Pohlman. Viewing stemming as recall enhancement. In *Proceedings of ACM SIGIR*, pages 40–48, 1996.

[13] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics - Doklady 10, vol. 10*, pages 707–710, 1966.

[14] X. Li, R. Singh, and R. Stern. Lattice combination for improved speech recognition. In *Proceedings of ICSLP*. CMU, 2002.

[15] R. Ordelman, A. van Hessen, and F. de Jong. Compound decomposition in dutch large vocabulary speech recognition. In *Eurospeech 2003*, 2003.

[16] M. Porter. An algorithm for suffix stripping. *Program 14(3)*, pages 130–137, 1980.

[17] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[18] P. Schauble. *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, 1997.

[19] M. Siegler. Integration of continuous speech recognition and information retrieval for mutually optimal performance. In *PhD Thesis*. CMU, 1999.

[20] L. van der Werff, W. Heeren, R. Ordelman, and F. de Jong. Radio oranje: Enhanced access to a historical spoken word collection. In *Proceedings of CLIN17*, 2007.

[21] E. M. Voorhees. Overview of the sixth text retrieval conference (trec-6). Department of Commerce, National Institute of Standards and Technology, 1997.

[1] http://hmi.ewi.utwente.nl/choral

[2] http://www.nwo.nl/catch