INFORMATION FUSION FOR SPOKEN DOCUMENT RETRIEVAL

Kenney Ng

Spoken Language Systems Group
MIT Laboratory for Computer Science
545 Technology Square, Cambridge, MA 02139 USA

ABSTRACT

In this paper we investigate the fusion of different information sources with the goal of improving performance on spoken document retrieval (SDR) tasks. In particular, we explore the use of multiple transcriptions from different automatic speech recognizers, the combination of different types of subword unit indexing terms, and the combination of word and subword-based units. To perform retrieval, we use a novel probabilistic information retrieval model which retrieves documents based on maximum likelihood ratio scores. Experiments on the 1998 TREC-7 SDR task show that the use of these different information fusion approaches can result in significantly improved retrieval performance.

1. INTRODUCTION

Spoken document retrieval (SDR) is the task of searching a static collection of recorded speech messages in response to a user-specified natural language text query and returning an ordered list of messages ranked according to their relevance to the query. The development of automatic methods to index, organize, and retrieve spoken documents will become more important as the amount of spoken language data continues to grow. At the same time, the development of these methods will have a significant impact on the use of speech as a data type because speech is currently a difficult medium to browse and search efficiently. In this work, we investigate the use of information fusion to try to improve SDR performance. In particular, we explore the use of multiple transcriptions from different automatic speech recognizers, the combination of different types of subword unit indexing terms, and the joint use of word and subword units. Experiments on the 1998 TREC-7 SDR task show that a number of different information fusion approaches can significantly improve retrieval performance.

The paper is organized as follows. We first present our novel probabilistic information retrieval model in Section 2 and describe the data corpus that comprise the TREC-7 SDR task in Section 3. We then present a series of retrieval experiments. In Section 4, we establish a reference retrieval performance level using manually created word transcriptions of the spoken documents and then examine the use of word transcriptions generated by several large vocabulary automatic speech recognition (ASR) systems. The output from the different ASR systems are examined individually and in combination. We then explore, in Section 5, the use of subword unit indexing terms derived from the manual and ASR transcriptions and measure their ability to perform retrieval. Different subword units are examined separately and in combination. Finally, we investigate the effect of combining both word and subword units in Section 6 and close with some conclusions in Section 7.

2. INFORMATION RETRIEVAL MODEL

Given a collection of n documents, $\{D_i\}_{i=1}^{n}$, each document D_i has a prior likelihood given by $\mathrm{p}(D_i)$. After a query Q is specified by a user, the likelihood of each document changes and becomes that given by the conditional probability: $\mathrm{p}(D_i|Q)$. Some documents will become more likely after the query is specified while others will either remain the same or become less likely. The documents that become more likely are probably more useful to the user and should score better and be ranked ahead of those that either stay the same or become less likely. As a result, we propose to use the relative change in the document likelihoods, expressed as the likelihood ratio of the conditional and prior probabilities, as the metric for scoring and ranking the documents:

$$\mathrm{S}(D_i, Q) = \mathrm{p}(D_i|Q) \,/\, \mathrm{p}(D_i) \tag{1}$$

We can decompose this likelihood ratio score into more easily estimated components using Bayes' Rule and rewrite (1) as:

$$\mathrm{S}(D_i, Q) = \frac{\mathrm{p}(Q|D_i)\,\mathrm{p}(D_i) \,/\, \mathrm{p}(Q)}{\mathrm{p}(D_i)} = \frac{\mathrm{p}(Q|D_i)}{\mathrm{p}(Q)} \tag{2}$$

where $\mathrm{p}(Q|D_i)$ is the probability of query Q given document D_i and $\mathrm{p}(Q)$ is the prior probability of query Q. We assume that the query imposes a multinomial distribution over the set of possible terms in the corpus so that $\mathrm{p}(Q|D_i)$ and $\mathrm{p}(Q)$ can be modeled as a product of term probabilities, $\mathrm{p}(t|D_i)$ and $\mathrm{p}(t)$, over the terms t in query Q. To address the sparse training data issue, we use Good-Turing methods to estimate $\mathrm{p}(t)$ and a back-off mixture model to estimate $\mathrm{p}(t|D_i)$. The resulting retrieval score measure is:

$$\mathrm{S}(D_i, Q) = \prod_{t \in Q} \left(\frac{\alpha\,\mathrm{p}(t|D_i) + (1-\alpha)\,\mathrm{p}(t)}{\mathrm{p}(t)} \right)^{q(t)} \tag{3}$$

where $q(t)$ is the number times term t occurs in query Q, and α is the back-off mixture weight which is estimated automatically using the EM algorithm to maximize the likelihood of the query.

Automatic relevance feedback is a well-established method for improving retrieval performance [1]. It works by running a second retrieval pass using a query constructed by modifying the original query using information from the top scoring documents obtained from a preliminary retrieval pass. We extend our basic retrieval model to include an automatic feedback processing stage by developing a new query reformulation algorithm that is specific to our probabilistic model. The objective of the algorithm is to increase the likelihood ratio score of a joint document composed of the top-ranked documents from the preliminary retrieval pass. This is done by removing certain terms from the original query and adding new terms from the top-ranked documents with appropriate term weights. The idea is that improving the document scores should lead to better retrieval performance. A complete description of our probabilistic retrieval model can be found in [2].

No. of documents	2866		
No. of topics	23		
	Min.	**Mean**	**Max.**
Document length (words)	2	269	12594
Topic length (words)	5	14.7	27
No. of relevant docs/topic	1	17	60

Table 1: Statistics for the TREC-7 SDR data set.

```
<Section S_time=1881.464 E_time=1896.492 ID=eh971009.37>
 defense secretary william cohen has issued a fresh
 warning to iraq about violating the air exclusion
 zones in the north and south of the country mr cohen
 said iraqi pilots would have to bear the consequences
 if they continue to violate the zones
</Section>
```

Figure 1: A sample TREC-7 SDR document (reference transcript).

3. DATA CORPUS

Experiments are done on the spoken document retrieval (SDR) task from the 1998 Text REtreival Conference (TREC) sponsored by the National Institute of Standards and Technology (NIST) [1]. Statistics for the TREC-7 SDR data set are shown in Table 1.

The document collection consists of 2866 news stories drawn from approximately 100 hours of recorded radio and television news broadcasts from the following information sources: *ABC World News Tonight*, *CNN Early Prime*, *CNN Headline News*, *CNN Primetime News*, *CNN The World Today*, *C-SPAN Public Policy*, *C-SPAN Washington Journal*, and *PRI The World*. A sample document (reference transcription) is shown in Figure 1.

There are 23 queries (also called "topics"), numbered 51-73, in this test set. Each topic consists of a natural language text sentence describing an information request. A sample topic (number 68) is shown in Figure 2. To evaluate the performance of an information retrieval system, the retrieved messages are evaluated against relevance assessments created for each topic. Basically, all the relevant documents in the collection needs to be identified for each of the topics. There is a total of 390 relevant documents for the 23 topics. From Table 1, we see that the number of relevant documents for each topic can vary greatly: some topics have many relevant documents while others only have a few.

Retrieval performance is measured in terms of a tradeoff between *precision* and *recall*. Precision is the number of relevant documents retrieved over the total number of documents retrieved. Recall is the number of relevant documents retrieved over the total number of relevant documents in the collection. Because it is sometimes difficult to compare the performance of different retrieval systems using precision-recall curves, a single number performance measure called *mean average precision* (mAP) is commonly used [1]. It is computed by averaging the precision values at the recall points of all relevant documents for each query and then averaging those across all the queries in the test set. In this paper, we report retrieval performance using this mAP metric.

```
What are confirmed incidents of U.S. military air
crashes and what types of aircraft were involved?
```

Figure 2: A sample TREC-7 SDR query/topic (number 68).

System Description	mAP	
	Preliminary	**Feedback**
No Preprocessing	0.3423	0.3850
Stop word removal	0.3348	0.4307
Word stemming	0.4143	0.4521
Stopping + stemming	0.4256	**0.5295**

Table 2: Retrieval performance in mean average precision (mAP) on the TREC-7 SDR task using reference word transcriptions.

4. WORD-BASED INDEXING TERMS

4.1. Reference Transcriptions

We first establish a reference retrieval performance by using manually generated word transcriptions of the spoken documents. We assume that there are no errors in these reference transcriptions. The resulting retrieval performance can be considered the upper bound performance since this is equivalent to using the output of a perfect speech recognizer.

Table 2 shows retrieval performance in mean average precision (mAP) on the TREC-7 SDR task using reference word transcriptions of the spoken documents. Performance after the preliminary retrieval pass and the second automatic feedback pass are shown. Performance is also broken down to indicate the effect of stop word removal using a fixed list of 220 common English function words and the effect of using a standard (Porter's) stemming algorithm to conflate the words. The automatic feedback process significantly improves retrieval performance in all cases. Each preprocessing step also helps improve performance and the use of both together results in additive gains. The final retrieval performance is mAP = 0.529 which is competitive with the performance of the top retrieval systems reported in TREC-7 [1].

4.2. ASR Transcriptions

Since current state-of-the-art speech recognizers are not perfect, we next measure retrieval performance using word transcriptions of the spoken documents generated automatically by large vocabulary continuous speech recognition systems. Eight sets of transcriptions with varying word error rates generated by different speech recognizers are examined. These transcriptions were generated by the sites participating in the TREC-7 SDR task and are distributed along with the reference transcriptions by NIST to facilitate and encourage "cross-recognizer" experiments [1].

Retrieval performance, in mean average precision (mAP), as a function of speech recognition performance, in word error rate (WER), on the TREC-7 SDR task is shown in Table 4 under the column labeled "word." Retrieval performance is highly correlated with recognition performance: the lower the WER, the better the mAP. This can be seen more clearly in the mAP versus WER curve ($\triangle$) labeled "word" plotted in Figure 5. Even though the recognition error rates are relatively high (the best WER is 24.6%), retrieval performance using these errorful ASR transcriptions is only about 5% worse than using the error-free reference transcriptions (mAP of 0.5025 vs. 0.5295). This indicates that many of the words useful for retrieval were correctly recognized.

4.3. Combining Multiple ASR Outputs

The ASR transcriptions used in the experiments above consist of only the single most likely word sequence hypothesized by the speech recognizer. No information about the confidence or likelihood of each recognized word is available. However, by combining the outputs of several different recognizers, we can compute a

rough estimate of the term occurrence probabilities for each recognized word. For example, we can use a simple maximum likelihood estimate based on the number of occurrences of the word in the combined set of recognition hypotheses:

$$\sum_r \mathrm{c}(t|D_i^r) \;/\; \sum_{r,t} \mathrm{c}(t|D_i^r) \qquad (4)$$

where $\mathrm{c}(t|D_i^r)$ is the number of occurrences of term t in the transcription of document D_i generated by recognizer r. This additional information can be directly incorporated into our probabilistic retrieval model in the $\mathrm{p}(t|D_i)$ and $\mathrm{p}(t)$ terms to better reflect the quality of the ASR transcriptions. Using this term occurrence probability estimate leads to improved retrieval performance as shown in the rows labeled "combined" in Table 4. We examine the use of all eight ASR outputs (All 8) and the use of only the top five (Top 5) to estimate the term occurrence probabilities. Performance using information derived from combining all eight systems is better than the performance of the single best system (mAP of 0.5073 vs. 0.5025). Restricting the combination to systems with the best speech recognition performance (Top 5) gives even better results (mAP = 0.5125).

5. SUBWORD UNIT INDEXING TERMS

We now examine the use of subword units for information retrieval. The use of subword units in the recognizer constrains the size of the vocabulary needed to cover the language; and the use of subword units as indexing terms allows for the detection of new user-specified query terms during retrieval [3]. Subword units consisting of overlapping, fixed-length, phone sequences ranging from n=2 to n=5 in length with a phone inventory of 41 classes are used. These subword units are derived by successively concatenating the appropriate number of phones from phonetic transcriptions of the spoken documents. Examples of n-phone subword units (n=1, 2) for the phrase "weather forecast" are shown in Table 3.

Phonetic transcriptions of the spoken documents are obtained from the reference and ASR word transcriptions by mapping the words to their corresponding phone strings using a pronunciation dictionary. Word and sentence boundary information is lost during this process resulting in each document being treated as a single long phone sequence. The subword units are then created from these phonetic transcriptions. Similar processing is done on the topics to convert them to matched subword unit representations. We note that phonetic transcriptions obtained in this way are suboptimal since cross-word coarticulation effects are not captured.

5.1. Reference Transcriptions

Retrieval performance, in mAP, on the TREC-7 SDR task using a range of n-phone subword units ($n = 2, \ldots, 5$) derived from the reference document transcriptions is shown in Figure 3. Performance is shown for the baseline system (□), the use of stop term removal (△), and the use of normalized (stopped and stemmed) topic descriptions (○). We note that subword units of intermediate length (n=3,4) perform better than short (n=2) or long (n=5) units in all three cases. This is due to a better tradeoff of the intermediate length units between being too short and matching too many terms and being too long and not matching enough terms.

Subword Unit	Indexing Terms
word	weather forecast
1phn (n=1)	w eh dh er f ow r k ae s t
2phn (n=2)	w_eh eh_dh dh_er er_f f_ow ow_r r_k k_ae ae_s s_t

Table 3: Examples of n-phone subword unit indexing terms.

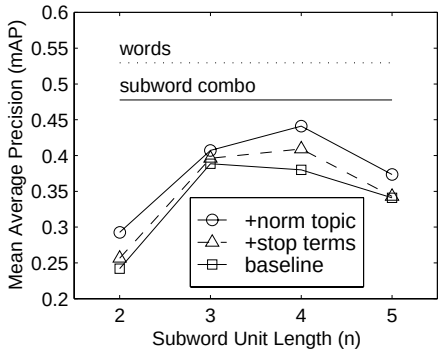

Figure 3: Retrieval performance (mAP) using a range of n-phone subword units derived from the reference document transcriptions.

In Section 4, we saw that the removal of frequently occurring non-content "stop" words helped retrieval performance. To determine if similar processing can improve performance for subword units, we explore the use of "stop term" removal. For each type of subword unit (i.e., different n), a set of stop terms is created by computing the document frequency of each term, rank ordering them in decreasing order, and then thresholding the list to select terms that occur in a large fraction of the documents (>25%) in the collection. The number of stop terms varies depending on the type of subword unit and ranges from a low of 28 for n=2 to a high of 600 for n=3. The resulting stop terms consist mainly of short function words and common prefixes and suffixes. As shown in Figure 3, the use of stop term removal results in small but consistent improvements for all the different subword units (△).

Unlike the topic descriptions used in the word-based experiments described in Section 4, the topics used in the subword experiments are not normalized using stop word removal and word stemming. Since normalization helped with word units, we wanted to explore its effect on subword units. We start with the normalized word-based topic descriptions, convert the words to their phonetic representation using a pronunciation dictionary, and then generate the n-phone subword units to obtain the subword-based topics description. As shown in Figure 3, performance for all subword units is significantly improved when using these normalized topics (○).

5.2. Combining Multiple Subword Units

Different subword unit representations can capture different types of information. For example, longer subword units can capture word or phrase information while shorter units can model word fragments. The tradeoff is that the shorter units are more robust to errors and word variants than the longer units but the longer units capture more discrimination information. One simple way to try to combine the different information is to form a new document-query retrieval score by linearly combining the individual retrieval scores obtained from the separate subword units:

$$\mathrm{S}'(D_i, Q) = \sum_n w_n \; S^n(D_i, Q) \qquad (5)$$

where $S^n(D_i, Q)$ is the normalized (zero-mean and unit-variance) document-query score (3) obtained using subword representation n and w_n is a tunable weight parameter. Empirically determined combination weights of w_2 = w_5 = 0.1 and w_3 = w_4 = 1.0 are used. Performance using the combined subword units is shown as the solid line in Figure 3. Performance is significantly better than that of the best individual subword unit (n=4): mAP of 0.4776 vs. 0.4411. However, it is still not as good as using the reference word transcriptions: mAP = 0.5295 (dotted line).

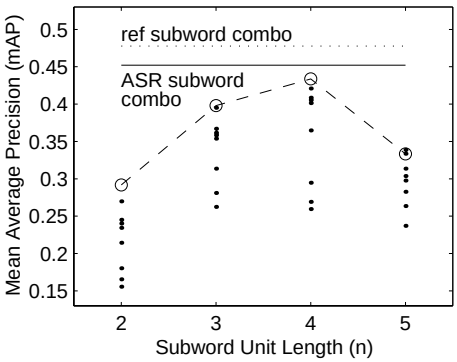

Figure 4: Retrieval performance (mAP) using a range of n-phone subword units derived from speech recognition transcriptions.

5.3. ASR Transcriptions

Retrieval performance (mAP) using a range of n-phone subword units derived from individual ASR transcriptions is plotted (•) in Figure 4. The performance variability is due to the different error rates of the ASR systems. For each ASR system, improved performance can be obtained by combining the different subword units using (5). Retrieval performance using combined subword units for each ASR system is shown in Table 4 under the column labeled "subword" and plotted (□) in Figure 5. Performance using subword units is consistently about 0.06 mAP points worse than using word-based units (△), indicating some loss of information.

5.4. Combining Multiple ASR Outputs and Subword Units

Previously, we saw that combining multiple ASR outputs to estimate term occurrence probabilities using (4) and combining the different subword units using (5) each improves retrieval performance. We now explore using both of these combinations together. Figure 4 shows the results of doing this two-stage combination. We start with a range of n-phone subword units derived from individual ASR transcriptions (•). Next, we combine the outputs from the top 5 ASR systems for each subword unit type using (4) to estimate term occurrence probabilities. This results in performance (◯) that is slightly better than the best individual system. Finally, we combine the different subword units using (5) which further improves performance to mAP = 0.4522 (solid line). This is also the performance shown in the row labeled "combined" in Table 4. As a reference, performance using combined subword units derived from clean transcripts is mAP = 0.4776 (dotted line).

6. COMBINING WORD AND SUBWORD UNITS

Finally, we examine the combination of word and (combined) subword units to see if this information fusion can result in improved retrieval performance. We use (5) to combine the results of the two different types of indexing terms with equal (w_n = 1.0) weights. The performance of the word and subword combination is shown in Table 4 under the column labeled "sub+word" and plotted (◯) in Figure 5. Performance using the combined units is consistently better than using just the word (△) or subword units (□) alone. Performance using ASR transcriptions (combined top 5) improves to mAP = 0.5389! Even with clean reference transcriptions, performance using the combined word and subword units is significantly better than using just words alone (mAP of 0.5564 vs. 0.5295). Analysis indicates that the subword units add the flexibility of partial word matches and an increased discrimination capability of cross-word constraints to the word units.

System Description	WER (%)	Mean Average Precision (mAP)		
		Word	Subword	Sub+Word
Individual ASR Systems	66.0	0.3382	0.2907	0.3464
	61.3	0.3812	0.3157	0.3784
	46.6	0.3357	0.2716	0.3390
	35.6	0.4681	0.3756	0.4758
	33.8	0.4779	0.4200	0.5065
	31.0	0.4904	0.4266	0.5172
	29.5	0.4705	0.4135	0.5084
	24.6	0.5025	0.4463	0.5243
Combined (All 8)	–	0.5073	0.4320	0.5304
Combined (Top 5)	–	0.5125	0.4522	0.5389
Reference	0.0	0.5295	0.4776	0.5564

Table 4: Retrieval performance (mAP) using subword, word, and combined (sub+word) indexing terms for individual ASR, combined ASR, and reference spoken document transcriptions.

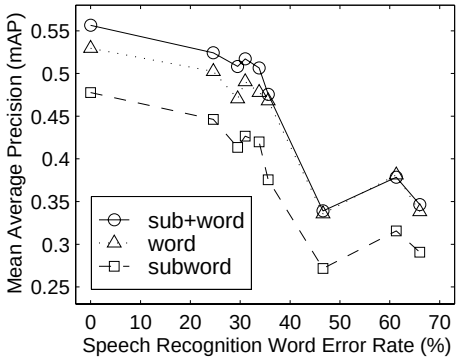

Figure 5: Retrieval (mAP) vs. speech recognition (WER) performance using subword, word, and subword+word indexing terms.

7. CONCLUSIONS

In this work, we investigate the fusion of different information sources with the goal of improving performance on spoken document retrieval tasks. We use a novel probabilistic information retrieval model and conduct experiments on the 1998 TREC-7 SDR task. We find that a number of different information fusion approaches can significantly improve retrieval performance: using multiple transcriptions from different speech recognizers to estimate term occurrence probabilities leads to improved performance; combining different types of subword unit indexing terms results in performance that is better than the best individual subword unit; and combining word and subword units improves performance over using just subword or word units alone.

8. ACKNOWLEDGMENTS

This research was supervised by Dr. Victor Zue and was supported by DARPA under contract N66001-96-C-8526, monitored through NCCOSC. We would like to thank the NLPIR and SNLP groups at NIST for making the TREC-7 SDR data available to us.

9. REFERENCES

[1] D. K. Harman, ed., *Seventh Text REtrieval Conference (TREC-7)*, NIST-SP 500-242, 1998.

[2] K. Ng, "A maximum likelihood ratio information retrieval model," in *Proc. Eighth Text REtrieval Conference (TREC-8)*, 1999.

[3] K. Ng and V. Zue, "Subword unit representations for spoken document retrieval," in *Proc. Eurospeech '97*, pp. 1607–1610, 1997.