Experiments in Spoken Document Retrieval at CMU

M. Siegler, A. Berger, M. Witbrock, A. Hauptmann*
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

*Justsystem Pittsburgh Research Center
4616 Henry St.
Pittsburgh, PA 15213

Abstract

We describe our submission to the TREC-7 Spoken Document Retrieval (SDR) track and the speech recognition and the information retrieval engines. We present SDR evaluation results and a brief analysis. A few developments are also described in greater detail including:

- A new, probabilistic retrieval engine based on language models.
- A new, TFIDF-based weighting function that incorporates word error probability.
- The use of a simple confidence estimate for word probability based on speech recognition lattices.

Although improvements over a development test set were promising, the new techniques failed to yield significant gains in the evaluation test set.

1. The SDR Data and Task

The entire set of speech data for the 1998 TREC-7 spoken document retrieval track consisted of 153 hours of broadcast news, approximately 80 for training and 73 for testing. The data had been segmented into stories and manually transcribed. In the test set, there were three "versions" of the data available: A manually generated transcript, speech recognition transcripts based on IBM and CMU recognizers, and the raw audio data, to be transcribed by our own recognizer.

The entire training set was used to train acoustic models for the speech recognition system. The remainder was held out as unseen test data. There were about 3245 stories in the training data set and 2866 in the test set. To develop and debug the system, the TREC-6 evaluation set was used in a *Known-Item Retrieval* system -- where every query has only one document assigned as relevant.

In our experiments on the evaluation test set, the average precision of the retrieval for each of the relevant documents was used to judge the quality of the retrieval. However, since relevance judgements were not available for the development test set, we used last year's *Average Inverse Rank* to judge retrieval quality.

2. System Overview

In this section we give a system description of the actual CMU TREC-7 SDR submission. The speech recognition system is outlined as well as a fully automatic information retrieval weighting scheme suitable for retrieving documents transcribed (with errors) by automatic speech recognition.

The Speech Recognition Component

The Sphinx-III speech recognition system used for this evaluation was configured similar to the 1997 TREC-7 SDR evaluation [1], although several changes have been made since then. Sphinx-III is a large vocabulary, speaker independent, fully continuous hidden Markov model speech recognizer with separately trained acoustic, language and lexical models.

For the current evaluation a gender-independent HMM with 6000 senonically-tied states and 16 diagonal-covariance Gaussian mixtures was trained on the TREC-7 SDR training set.

The decoder used a Katz-smoothed trigram language model trained on the 1992-1996 Broadcast News Language Modeling (BN LM) corpus and the LDC-provided supplemental newswire (NW) data from 1997-1998. This is a fairly standard language model, much like those that have been used in the DARPA speech recognition community for the past several years. The lexicon was chosen from the most common words in this corpus. For this evaluation, the vocabulary was comprised of the most frequent 64k words in the BN LM + NW corpora.

The Information Retrieval Component

Both documents and queries were processed using the same conditioning tools, namely noise filtering, stopword removal, and term stemming:

- Noise Filtering: The goal of noise filtering was simply to remove non-alphabet ASCII characters, punctuation, and other junk considered irrelevant to IR. All punctuation was removed except for spelled-letter words, e.g. "C.M.U," and the use of the apostrophe for contractions, e.g. "CAN'T." Any changes in case were removed.
- Stopword removal: A set of 811 stopwords was compiled from a combination of the SMART IR engine and several selected by hand based on document frequency. These words were removed entirely.
- Term mappings: A set of 4578 mappings was used to map words with irregular word endings that were not properly covered by an implementation of the Porter algorithm. An on-line Houghton-Mifflin dictionary was used for this lookup of irregular words and their roots.
  An example of this mapping is `APPENDICES`→`APPENDIX`
- Term stemming: An implementation of the Porter algorithm was applied to map words to their common root.

For this evaluation, we had two different relevance weighting schemes using entirely different approaches. The first was a vector-space model built on the LNU weighting scheme [3], whereas the second used a language model approach to estimate likelihoods.

3. Word Probability In The Relevance Equation: Mutual Information

It is some combination of the two factors frequency and selectivity that is used to evaluate the relevance of documents to queries. Many retrieval engines use derivatives of Salton's vector space model, specifically a measure commonly known as TFIDF (Term Frequency by (log) Inverse Document Frequency.)

Given a set of M documents, a word w_i, and a specific document D_m, the IDF is defined as:

$$IDF_i \equiv -\log\left(\frac{\left|\{\forall m \text{ s.t. } w_i \in D_m\}\right|}{M}\right)$$

Although it is obvious that the IDF provides some measure of term selectivity, it is important, for its application in this paper, to derive a theoretical basis for its use. If documents and queries are regarded from a probabilistic point of view, the significance of IDF is readily apparent and motivates the use of word probabilities derived from the speech recognition.

Let documents and queries be defined as mappings of words into probabilities:

$$D : w_i \rightarrow P(w_i)$$
$$Q : w_i \rightarrow P(w_i)$$

The space of independent documents is defined as:

$$\boldsymbol{D} \equiv \{D_1, D_2, \cdots, D_M\}$$

The *a-priori* probabilities of document relevance are equal:

$$P(D_m) = 1/M$$

The probability of a document, given a particular word is:

$$P(D_m \mid w_i) = P(w_i \mid D_m)\frac{P(D_m)}{P(w_i)}$$

And by simple expansion:

$$P(D_m \mid w_i) = \frac{P(w_i \mid D_m)P(D_m)}{\sum_{m'=1}^{M} P(w_i \mid D_{m'})P(D_{m'})}$$

Consider the information content of word w_i to be the mutual information of the document set and the word:

$$I(\boldsymbol{D}; w_i) \equiv H(\boldsymbol{D}) - H(\boldsymbol{D} \mid w_i)$$

Expanding, using the definition of entropy:

$$I(\boldsymbol{D}; w_i) = -\sum_{m=1}^{M} P(D_m)\log_2 P(D_m) + \sum_{m=1}^{M} P(D_m|w_i)\log_2 P(D_m|w_i)$$

The relevance of query Q to document D_m in space $\boldsymbol{D}$ is defined as the expected value of this information content:

$$\mathrm{Rel}(Q, D_m|\boldsymbol{D}) = \sum_{i=i}^{N} E\{I(\boldsymbol{D}; w_i)|Q, D_m)\}$$
$$= \sum_{i=1}^{N} P(w_i|Q, D_m)\, I(\boldsymbol{D}; w_i)$$

Assuming documents and queries to be independent:

$$\mathrm{Rel}(Q, D_m|\boldsymbol{D}) = \sum_{i=1}^{N} P(w_i|Q)P(w_i|D_m)I(\boldsymbol{D}; w_i) \qquad (1)$$

If documents and queries map words to indicator functions:

$$I(\boldsymbol{D}; w_i) = \log_2(M) - \sum_{i=1}^{N} 1(w_i|D_m) = \mathrm{idf}(w_i|\boldsymbol{D})$$

The the relevance function reduces to the familiar:

$$\mathrm{Rel}(Q, D_m|\boldsymbol{D}) = \sum_{i=1}^{N} 1(w_i|D_m)1(w_i|Q)\,\mathrm{idf}(w_i|\boldsymbol{D})$$

Where the indicator functions are the TF values. By this logic, it is seen that the IDF can be supported as a meaningful derivative of information content. In addition, the more general form in Equation 1 can be used when word probabilities are available.

4. Estimating Word Probability From Recognition Lattices

In the process of decoding the incoming speech signal into a word string, the Sphinx III recognizer produces a lattice of words representing the many competing hypotheses. Each hypothesized word in the lattice has a starting time, an ending time, a link to possible following words, and model probabilities for this word. After producing this lattice, the recognizer selects the most probable path after weighing evidence from the different modeling sources available. The best path, also called the *top-1* hypothesis, is generated as the output of the recognizer.

Although the lattice is available, only the best path has typically been used for the purpose of information retrieval. Although the lattice does not contain all possible word sequences, it is a far more detailed representation of what may have been said than can be given in a single transcription. One serendipitous benefit of the lattice is that the presence of a large number of options at any moment in time may indicate an uncertainty in word recognition. This is valuable, since it would be beneficial to predict which words in the top-1 hypothesis are incorrect, and discount them during information retrieval.

One way of measuring the number of competing hypotheses for a specific node in a lattice is the following:

- Count the time span (in frames) of the node: N
- Count the number of frames contained in other nodes that occur simultaneously with this node (partially or completely): M
- The Lattice Occupation Density (LOD) is N/(N+M).

In the example shown in Figure 1, the recognition system is less certain about the presence of “today” than “news" because the latter word has no competing hypotheses.

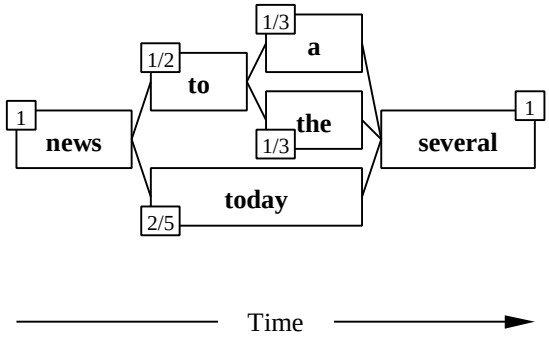

Figure 1: A simple lattice. Numbers show the
Lattice Occupation Density (LOD) values for the various nodes

The TREC-6 training corpus was used to build a probability model by analyzing the lattices created during recognition. The LOD values for each word in the top-1 hypotheses were collected, and the word errors tallied. In Figure 2, the probability that a hypothesized word occurred in the reference transcript is compared with its LOD value from the lattice. To use the measurements of the training set, a model of word probability was derived. The model used was a best fitting exponential of the form:

$$P(w \mid LOD) \approx 1 - 0.2^{LOD}$$

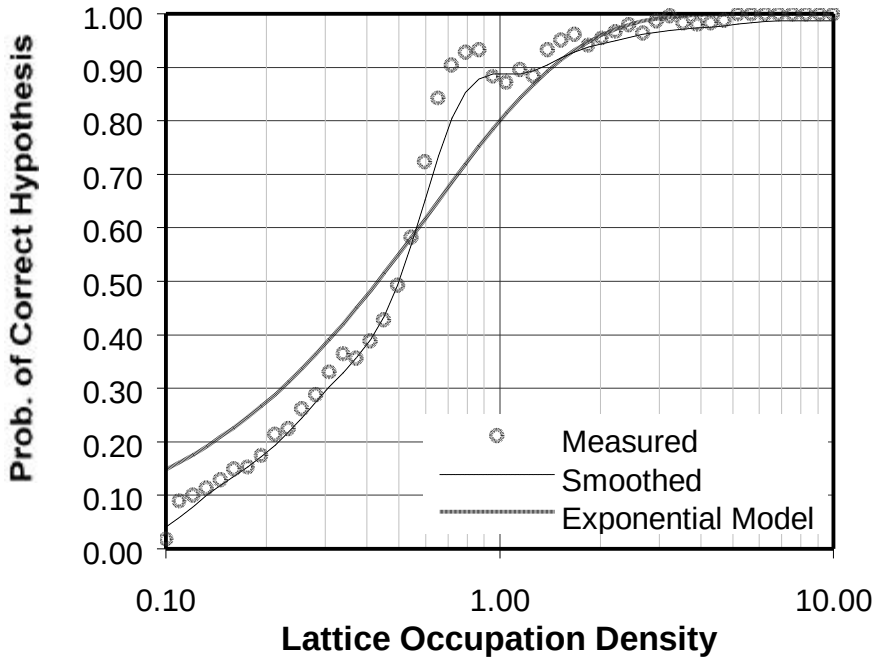

Figure 2: Using LOD to predict word probability in the top-1 hypothesis. For example, hypothesized words with an LOD of 1.0 appeared in the reference text approximately 85% of time.

Performance Of Mutual-Information Metric On Development Test Set

As can be seen in Table 1, using error prediction in conjunction with the mutual information metric can reduce the average inverse rank difference between reference texts and the top1 hypotheses by 25%.

Transcription Source	Average Inverse Rank	
	LNU	LNU+Mutual Inf.
REF	0.82	
TOP1	0.74	0.76

Table 1: Performance of the mutual information metric on the development test set.
Values are average inverse ranks for the different transcription sources.

5. Finding The Best IR Lattice Path

Although improving the word error rate of the transcription is the primary goal for speech recognition tasks, in the case of indexing and searching audio material, some consideration for the application of the transcription should be made. Since many types of "errors" in transcription are benign (for example a substitution of one stop-word for another) it is apparent that reducing the word error rate blindly does not necessarily translate to greater retrieval performance.

In order to clarify this point, the lattices from the speech recognition phase of the evaluation test set were scanned for their lowest word-error paths. This is commonly known as the *Oracle Word Error Rate* of the lattice. In addition to these paths, the paths that generate the lowest word error after text-processing *with respect to the text-processed references* were generated. The distinction of the latter is that paths containing errors in regions that are benign are not penalized for these errors.

Table 2 shows the performance of the oracle path selection using the mutual information retrieval engine on a 2601 document *subset* of the evaluation set. Approximately 10% of the original 2866 documents failed to produce lattices, or were too long to rescore, and these were deleted from the test set.

Filtering after the oracle path is found is tantamount to finding the path that generates the oracle word error rate of the lattice. As can be seen, this is not the optimal path through the lattice with respect to information retrieval. However, filtering before the optimal path is found is a better representative of an error criterion commensurate with the retrieval operation. What is surprising is that the original performance on the reference transcripts can be recovered from these lattice.

	Baseline		Oracle	
	Reference Transcripts	Speech Transcripts	Filter Before Oracle	Filter After Oracle
Training Set	AIR=0.79	AIR=0.75	AIR=0.79	AIR=0.75
Testing Set	P_{AVG}=0.39	P_{AVG} =0.36	P_{AVG} =0.39	P_{AVG} =0.37

Table 2: Baseline and Oracle Annotation on TREC-6 Training and Testing Sets.
Values are Average Inverse Rank (AIR) for the training set, and the average precision (P_{AVG}) for the testing set.

6. Information Retrieval Using Language Models

The approach to document retrieval described herein encodes every document into a succinct form, consisting of a set of index terms and a real-valued weight for each term. Loosely interpreted, the weight for a term corresponds to how much more likely the term is to appear in document than in the collection as a whole. For reasons which will become clear later, we refer to this compact representation as a language model for the document. We use penalized maximum likelihood to select, for each document, the optimal language model. In some sense, then, one can think of penalized maximum likelihood as a method for succinctly characterizing documents.

The first application of language modeling to IR was by Ponte and Croft, whose concern was document retrieval. The idea is to estimate a separate language model for each document in the collection, and then field queries by discovering the documents whose models accord most closely with the query. Estimating the parameters of these document-specific models is a delicate business, however. Maximum likelihood estimation---where the probability for a term is proportional to the number of times the term appeared in the document---won't work, since documents are often far too small for robust model estimation. The proposed solution is to interpolate or "smooth" each document-specific model with a model estimated on the entire collection of documents.

Table 3 shows the performance of the language model based retrieval engine on the development test set. In the general, the performance is on-par with the more finely tuned mutual information retrieval engine.

Transcription Source	Average Inverse Rank
REF	0.80
TOP1	0.76

Table 3: Performance of the language model based metric on the development test set. Values are average inverse ranks for the different transcription sources.

7. Official TREC-6 SDR Results

Table 4 shows the official CMU TREC SDR results. Unfortunately, the performance of the mutual information retrieval engine, coupled with the probability estimator did not fare as well as hoped. Upon investigation, the model using the lattice occupation density turned out to be a poor estimate of the word probability for the evaluation test set. However, the language model retrieval engine fared well in either case.

Transcription Source	LNU metric P_{AVG}	LM metric P_{AVG}
REF	0.36	0.39
B1	0.33	0.35
B2	0.26	0.27
CMU	0.32	N/A
CMU+Lattice	0.29	N/A

Table 4: Performance of the CMU TREC-7 SDR Evaluation System

References

[1] M. Siegler, M. Witbrock, S. Slattery, K. Seymore, R. Jones, A. Hauptmann, "Experiments in Spoken Document Retrieval at CMU," *Proceedings of TREC-6*, November 1998, Gaithersburg, MD.

[2] M. Siegler, M. Witbrock, "Improving The Suitability Of Imperfect Transcriptions For Information Retrieval From Spoken Documents," *ICASSP '99*, March 1999, to be published.

[3] A. Singhal, C. Buckley, M. Mitra, "Pivoted Document Length Normalization," *SIGIR-1996*, Zurich, Switzerland, August 1996.

Acknowledgments

This research was supported in part by DARPA under research contract F33615-93-1-1330 and N00039-91-C-0158. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DARPA or the U. S. Government.