IMPROVING THE SUITABILITY OF IMPERFECT TRANSCRIPTIONS FOR INFORMATION RETRIEVAL FROM SPOKEN DOCUMENTS

Matthew Siegler
msiegler@cs.cmu.edu
Dept. Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Avenue,
Pittsburgh PA 15213

Michael Witbrock
witbrock@justresearch.com
Justsystem Pittsburgh Research Center
4616 Henry St,
Pittsburgh PA 15213

ABSTRACT

Recently there has been a considerable focus on information retrieval for multimedia databases. When speech is used as the source material for multimedia indexing, the effect of transcriber error on retrieval effectiveness must be considered. This paper describes a method for measuring the relevance of documents to queries when information about the probability of word transcription error is available. To support the use of this technique, a method is presented for estimating word error probability in speech recognition engines that use word graphs (lattices). An information retrieval experiment using this technique on a large corpus of spoken documents is discussed. The method was able to reduce the difference in retrieval effectiveness between reference texts and hypothesized texts by 13%-38% depending on the size of the document set.

1. INTRODUCTION

Both the Computer Speech Recognition (CSR) and Information Retrieval (IR) components of a multimedia retrieval system such as Informedia [1] are imperfect methods for describing the information content of a spoken document and locating it based on a query. In addition, their technologies have been developed in relative isolation from one another, and their interactions have only been studied superficially. Consequently, there is a danger that an independent improvement in only the CSR or IR component will not improve overall system performance, and that it might, in some cases, actually *impair* performance.

In general, the IR systems that work with CSR-generated information tend to take a "black-box" approach; each system is designed, implemented, and optimized in the absence of the other. The result is that substantial performance gains in the CSR systems have not been reflected in the IR systems that are mated with them. This research describes an attempt to overcome the shortcomings of previous work by tying the retrieval engine more closely to the operation of the speech recognition system.

2. USING WORD PROBABILITY IN THE RELEVANCE EQUATION

In many text retrieval systems, the first step is to map the word space of the source documents into a smaller space. This is often carried out through the removal of a set of commonly occurring words [2], and by merging words sharing the same root into a new term [3]. Finally, the documents and queries are represented in a vector space model, with each word's count as an element of the vector [4].

When comparing a document with a query in a retrieval application, it is commonplace to compute a weighted inner product of the two. This inner product is provides a measure of relevance, and documents are selected based their high scores on this measure.

Typically, each word is given a weighting factor that reflects its relative selectivity for identifying particular documents. For example, a word that occurs in almost every sentence would not be seen as very selective, and would therefore be properly discounted during the relevance computation.

In addition, the number of times a word occurs in a document is indicative of its relative importance to that document.

It is some combination of these 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 [4], 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.

3. PREDICTING WORD ERRORS

In the process of decoding the incoming speech signal into a word string, the Sphinx III [5] 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 [6]. 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.

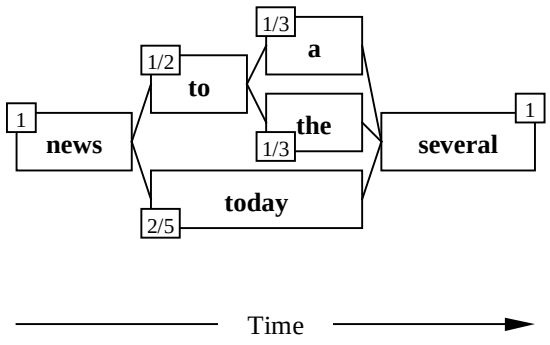

Figure 1: A simple lattice. Numbers show the Lattice Occupation Density (LOD) values for the various nodes.

4. EXPERIMENTS

4.1 Training and Testing Data

Speech data from the Spoken Document Retrieval (SDR) track of the 6th and 7th annual Text Retrieval Conferences (TREC) was used for training and testing the probability model and relevance equations [7][8]. Counting all training and testing material, there are approximately 153 hours of speech, in 6053 spoken documents, with a total of ~1.5 million words. The material consists of audio and transcripts for broadcast news articles during 1996-1997.

Three test sets were constructed to explore the effect of collection size on the retrieval methods. Table 1 shows the amount of data included in each set.

There were 49 queries in the test set, each specifically designed to select a single document from the corpus. This configuration is called a Known-Item Retrieval (KIR) task [9]. It is assumed, but not established, that all the remaining documents in the set are irrelevant to the query. Although KIR is a somewhat unrealistic retrieval scenario, it is easy to set up, and was used by NIST for the initial Spoken Document Retrieval (SDR) task at the TREC-6 conference. In evaluating the KIR task, the metric used is the *Average Inverse Rank* of the correct document.

Number of Documents	TREC-6 test	TREC-6 train	TREC-7 test
1421	yes	no	no
3187	yes	yes	no
6053	yes	yes	yes

Table 1: Data Sets used in the retrieval experiments, and the total number of documents they yielded.

4.2 System Configuration

The Sphinx-III speech recognition system was used for this experiment, in a similar configuration to that used in the 1997 DARPA BNT&UW evaluation [5]. Sphinx-III is a large vocabulary, speaker independent, fully continuous hidden Markov model speech recognizer with separately trained acoustic, language and lexical models. For this experiment, the decoder was run approximately ten times faster than normal[1], which resulted in a higher than usual error rate. In this configuration, the average word-error rate on broadcast news material is approximately 36%.

4.3 Deriving the Probability Model

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:

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

This model was applied to the top-1 hypothesis and the resulting estimated word probabilities were used in the information retrieval runs described below.

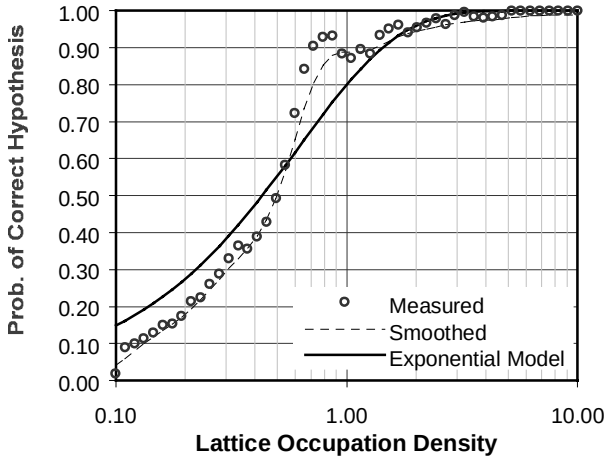

Figure 2: Using LOD to predict word probability in the top-1 hypothesis. For example, hypothesized words with an LOD of 1.0 appeared in the reference text approximately 85% of time.

[1] By reducing the beam used during acoustic search.

4.4 Information Retrieval Runs

The information retrieval system was run on each of the three, successively larger, document sets, and for each of three different text conditions, using the LNU relevance measure [9]. Table 2 and Figure 3 show that, by using the LOD metric, the degradation in retrieval performance for the recognized texts could be reduced by approximately 26% for the smallest, 38% for the middle sized, and by 13% for the largest document set.

Number of Documents	Reference	CSR Top-1	CSR Using LOD	Percent Gain
1421	0.82	0.74	0.76	26%
3187	0.68	0.60	0.63	38%
6053	0.65	0.57	0.58	13%

Table 2: Average inverse rank of the correct document in the information retrieval runs, and the performance gain by using the LOD metric and probability-weighted IR.

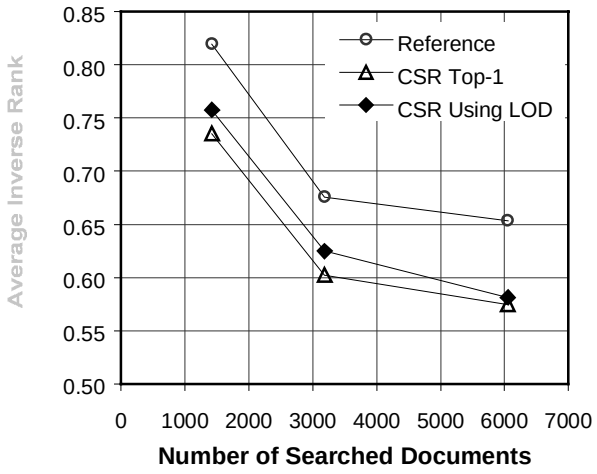

Figure 3: Chart of the data in Table 2. Each line shows the retrieval performance for a different source of text, under three different document sets.

5. DISCUSSION

The lattice occupation density metric introduced in this paper can be used to estimate word probabilities for the best recognition hypothesis, and this probability can be incorporated into the relevance equation to improve retrieval performance on spoken documents. For the 49 queries in a known-item retrieval task, the degradation for speech recognition texts in a set of 1421, 3817, and 6053 documents was reduced by 26%, 38% and 13% respectively. This is a satisfactory improvement, and justifies the use of probabilistic information derived from recognizer lattices in a well-motivated retrieval environment.

It is to be hoped that further improvements can be obtained by using additional information found in the lattice. Specifically, the system reported here made no use of candidate words other than those in the top-1 hypothesis, except as a means to cast doubt on the accuracy of their rivals. Consequently, the technique reported here could only help to ameliorate the effects of word insertions or substitutions of incorrect words for correct ones.

Retrieval accuracy is also affected by the absence of correct words that are skipped or that are replaced by incorrect words. Making use of the more words in the top-N hypotheses, in the probabilistic relevance framework presented here, may yield further significant improvements by enabling a future system to recover from this second class of error.

6. ACKNOWLEDGEMENTS

The authors would like to thank Alex Hauptmann, for his advisement in this work, and the CMU Speech group for patiently donating significant computational resources to the effort. This research was supported in part by DARPA under research contract F33615-93-1-1330 and N00039-91-C-0158. The preparation of this paper was supported in part by Justresearch, the Justsystem Pittsburgh Research Center.

7. REFERENCES

[1] Witbrock, Michael J. and Hauptmann, Alexander G. "Speech Recognition for a Digital Video Library", *JASIS: Journal of the American Society for Information Science*, 49(7), pp. 619-632, 1998.

[2] C. Rijsbergen, *Information Retrieval*, Butterworth, London, UK, 1975.

[3] M. Porter, "An algorithm for suffix stripping", *Program*, 14(3):130-137, July 1980.

[4] G. Salton, *The SMART retrieval system-experiments in automatic document processing*, Prentice-Hall, NJ, 1971.

[5] K. Seymore, S. Chen, S. J. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern and E. Thayer, "The CMU SPHINX III English Broadcast News Transcription System," *Proc. Broadcast News Transcription & Understanding Workshop*, Feb 1998.

[6] M. Siegler, M. Witbrock, S. Slattery, K. Seymore, R. Jones, and A. Hauptmann, "Experiments in Spoken Document Retrieval at CMU," *Proc. TREC-6*, 1997.

[7] E. Vorhees, D. Harman, "Overview of the Sixth Text REtrieval Conference," *Proc. of TREC-6*, Nov. 1997.

[8] "Overview of the Seventh Text REtrieval Conference," *Proc. of TREC-7*, Nov. 1998. (to appear)

[9] E. Vorhees, D. Harman, "Overview of the Fifth Text REtrieval Conference," *Proc. of TREC-5*, Nov. 1996.

[10] A. Singhal, C. Buckley, M. Mitra, "Pivoted document length normalization," *Proc. 19th ACM SIGIR*, 1996.