EXPERIMENTS IN SPOKEN QUERIES FOR DOCUMENT RETRIEVAL

J. Barnett[1], S. Anderson[1], J. Broglio[2], M. Singh[1], R. Hudson[3], and S. W. Kuo[3]

[1]Dragon Systems, 320 Nevada St., Newton, MA 02160 USA
Tel. (617) 965-5200, FAX: (617) 332-9575, E-mail: {stevea, monas}@dragonsys.com

[2]Sovereign Hill Software, 100 Venture Way, Hadley, MA 01035 USA
Tel. (413) 587-2222, FAX: (413) 587-2246, E-mail broglio@sovereign-hill.com

[3]Intermetrics Inc., 23 Fourth Ave., Burlington, MA 01803 USA
Tel. (617) 221-6990, FAX: (617) 221-6991, E-mail: {rgh, swk}@inmet.com

ABSTRACT

We report the results of three experiments using the errorful output of a large vocabulary continuous speech recognition (LVCSR) system as the input to a statistical information retrieval (IR) system. Our goal is to allow a user to speak, rather than type, query terms into an IR engine and still obtain relevant documents. The purpose of these experiments is to test whether IR systems are robust to errors in the query terms introduced by the speech recognizer. If the correctly recognized words in the search query outweigh the misinformation from the incorrectly recognized words, the relevant documents will still be retrieved. This paper presents evidence that speech-driven IR can be effective, although with a reduced precision. We also find that longer spoken queries produce higher precision retrieval than shorter queries.

For queries containing many (50-60) search terms and a recognizer word error rate (WER) of 27.9%, the precision at 30 documents retrieved is degraded by only 11.1%. For roughly the same WER, however, we find that queries shorter than 10-15 words suffer more than a 30% loss of precision.

1. INTRODUCTION

When using an information retrieval engine, a user typically types in a set of query words or phrases to retrieve relevant documents. However, for some applications (e.g., telephone-based retrieval, disabled users) speech would be the natural user interface. The first hurdle for a speech-driven IR is the degradation of retrieval performance due to errors in the query terms introduced by the speech recognition system. Because IR engines try to find documents that contain words that match those in the query, any errors in the query have the potential for derailing the retrieval of relevant documents.

IR systems retrieve documents based on statistical evidence from the word distributions in the input query and in the documents in the target collection. There are a wide variety of techniques in use for computing how well a given document matches a query, but most techniques involve some form of inverse-frequency term weighting that assigns more importance to less common words, since more common words, especially function words, have even distributions and therefore are not useful in distinguishing documents. If we enter the search terms by voice rather than as text, we introduce a number of errors in the query. However, we expect a certain robustness in the face of these recognition errors, because they often involve shorter function words, while content words are usually longer and easier to recognize.

Retrieving text documents using spoken queries has an inverse problem; retrieving audio and video documents that have been indexed using automatic speech recognition (ASR). Several authors ([3-8], [10-11]) have addressed this problem, and find very encouraging robustness in retrieval recall and precision in the face of errorful transcriptions of the indexed documents. Our task, however, is more challenging; the queries contain far fewer words than the documents to be retrieved, and so contain much less redundancy to overcome errors in the transcription.

We have done three experiments to measure the robustness of retrieval precision to 3 different levels of word error rate, for 4 different query lengths. The first, (prototype) experiment uses very long queries (50-60 words) and a single speaker, and demonstrates that retrieval precision can be very robust in the presence of recognition errors. The second experiment looks at correlations of precision loss with word error rate (WER), query length, and out-of-vocabulary (OOV) rate. The third experiment uses a much smaller database (of Boston Globe articles) and much shorter queries. It addresses the question of how retrieval precision degrades with shorter queries, which necessarily contain less redundant information.

2. RETRIEVAL PRECISION EXPERIMENTS

Experiments 1 and 2 use very long (50-60 word) queries to prove the principle that speech-based IR can be effective, while experiment 3 investigates the effect of using far fewer (2-15) query terms.

2.1. Experiment 1: Long Queries

The first experiment involves 35 queries from the TIPSTER set ([3]), which is a standard IR test collection. A single (male) speaker dictated the queries. Dragon's research LVCSR system was used for recognition with a 20K vocabulary and a bigram language model trained on the Wall Street Journal. By altering the width of the beam search, the mean word error rate (WER) was varied from 27.9% to 49.1% (The beam width was chosen because we believe that this yields relatively realistic recognition errors.) The resulting transcripts were used as queries to the INQUERY text retrieval system ([1]) and their results were compared with those of the text original. Table 1 shows the trade-off between WER and retrieval accuracy, with the column labeled '0%' representing the original text query.

The INQUERY system returns a list of documents, ranked in order of relevance to the query. Retrieval effectiveness is measured by considering the proportion of relevant documents ('precision') at different points in this list. Thus the leftmost column indicates that when the correct text of the queries was input to INQUERY, on average slightly more than 63% of the first 5 documents (and 60% of the first 30 documents, but only 27% of the first 500 documents) were relevant to the query.

	0%	27.9%	33.9%	49.1%
5 docs	0.634	0.646 *(+1.8%)*	0.663 *(+4.5%)*	0.651 *(+2.7%)*
30 docs	0.600	0.533 *(-11.1%)*	0.543 *(-9.5%)*	0.521 *(-13.2%)*
500 docs	0.265	0.238 *(-10.4%)*	0.238 *(-10.5%)*	0.227 *(-14.3%)*

TABLE 1: WER vs. Precision: TIPSTER queries

The last three columns show that precision is hurt by recognition errors, but that the deterioration is more sensitive to position in the ranked list than it is to the WER. Surprisingly, over the top 5 documents the errorful transcripts produce better results than the correct text. This result is presumably a fluke (in fact, experiment 2 shows a slight decline in precision at 5 documents), but the top of the ranking is apparently quite robust in the face of recognition errors. This is presumably because recognition errors are random from the point of view of topic. The highest ranked documents contain most or all of the query terms and the same documents are pulled to the top even when some of the terms are replaced by semantically random recognition errors. Further down the list, the documents contain fewer query terms, so that recognition errors have more of an effect. In no case, however, does the system fail catastrophically: even with a WER of 49%, the worst performance is a 14% relative loss (at 500 documents) or an 8% absolute loss (at 30 documents.)

2.2. EXPERIMENT 2: Correlations of Precision and WER.

This experiment investigates the correlation between recognition accuracy and retrieval accuracy. Using a slightly different version of the system than in experiment 1, we processed the same 35 queries by the same speaker. By way of comparison with the previous experiment, the mean WER was 25% and the precision at 5, 30, and 500 documents was 0.6286 (-0.9%), 0.5286 (-11.9%), 0.2330 (-12.2%), respectively. Note that in this case, performance at 5 documents is worse than the baseline, but only slightly.

As a measure of retrieval accuracy in this case, we took average precision over all relevant documents. (Average precision is calculated by computing the precisions at 10%, 20%, etc. recall rates and then averaging these precisions.) Average precision for the baseline was 0.3465, while for the spoken queries it was 0.3020 (-12.5%). The mean out-of-vocabulary (OOV) rate was 5.12%. The correlation between WER and loss of precision was 0.11, and the correlation between OOV rate and loss of precision was 0.14. The low correlation with WER is not surprising, since many of the errors involve function words which are ignored by the IR system. However, the lack of correlation with the OOV rate is striking, since OOV words are uncommon (not in the most frequent 20K words in the WSJ) and therefore are would be weighted heavily by the IR system. Stronger than either of these correlations, but still quite weak, is the correlation with query length, which is -0.18. We will examine the issue of query length more closely in the next experiment.

2.3. Experiment 3: Short Queries

There is one reason to view the results presented so far with caution, namely that the TIPSTER queries are very long, ranging in length from 20 to 165 words, with a mean length of 58 words. Thus, even with a 50% WER, enough good terms remain to retrieve the relevant documents.

In this third experiment we investigate shorter queries to see if retrieval precision is robust against recognition errors, as is true for long queries. We examine the loss of precision with queries consisting of 2-4, 5-8, and 10-15 content words. We call these "very short", "short", and "medium" length queries, to distinguish them from the longer TIPSTER queries.

The data set in this experiment consist of 24,630 Boston Globe articles published in the first six months of 1996. 10 queries of each of the 3 lengths ("very short", "short", and "medium" length) were constructed. The queries were chosen to have the following retrieval properties:

1) At least 2 of the top 5 articles relevant.
2) At least 5 of the top 20 articles relevant.
3) No more than 15 of the top 20 articles relevant.

Each article in the top 25 returned by entering the typed query was scored for relevance on a scale of 1-5, with 5 being very relevant and 1 being irrelevant. Articles with scores of "4" or "5" were marked as relevant for purposes of scoring precision.

15 speakers recorded each of the 30 queries (10 each of 3 lengths). There were 12 male and 3 female speakers.

Speech recognition was carried out with Dragon System's research LVCSR engine, using a speaker and gender independent (SI) acoustic model trained on 100 hours of speech (taken entirely from the Wall Street Journal corpus). The language model was built from three years of Boston Globe text, and contains 30,000 words. The queries were constructed by concatenating the top 5 recognition hypotheses; this was done in order to include correct query terms that may not have been in the recognizer's best transcription.

In addition to varying the query length, we repeated the recognition with three different parameter settings, each giving a different average word error rate. As in experiment 1, we varied a parameter that controls the beam width search, in order to produce realistic, but degraded, recognition performance.

The results for the "very short" queries (2-4 content words) are shown in Table 2 below. The results for the "short" and "medium" length queries are shown in Tables 3 and 4, respectively. In each table, the rows present retrieval precision for 5, 10, and 15 returned documents. The four columns represent the variation with word error rate, ranging from 0% (queries entered as text, without errors) to 50.8% word errors introduced by the speech engine.

	0% WER	30.0% WER	35.5% WER	50.8% WER
5 docs	0.780	0.605 *(-22%)*	0.508 *(-35%)*	0.344 *(-56%)*
10 docs	0.670	0.444 *(-34%)*	0.384 *(-43%)*	0.265 *(-60%)*
15 docs	0.553	0.343 *(-38%)*	0.299 *(-46%)*	0.214 *(-61%)*

TABLE 2: Precision (and relative loss of precision): "Very Short" queries**.**

	0% WER	30.0% WER	35.5% WER	50.8% WER
5 docs	0.920	0.675 *(-27%)*	0.576 *(-37%)*	0.409 *(-56%)*
10 docs	0.770	0.519 *(-33%)*	0.458 *(-41%)*	0.328 *(-57%)*
15 docs	0.593	0.405 *(-32%)*	0.364 *(-39%)*	0.262 *(-57%)*

TABLE 3: Precision (and relative loss of precision): "Short" queries.

	0% WER	30.0% WER	35.5% WER	50.8% WER
5 docs	0.760	0.592 *(-22%)*	0.527 *(-31%)*	0.441 *(-42%)*
10 docs	0.640	0.464 *(-28%)*	0.406 *(-37%)*	0.321 *(-50%)*
15 docs	0.527	0.348 *(-34%)*	0.309 *(-41%)*	0.249 *(-53%)*

TABLE 4: Precision (and relative loss of precision): "Medium" length queries.

3. DISCUSSION

Tables 2-4 show two trends that may be different aspects of the same result: more redundancy in the query increases IR robustness to errors in the query terms. The two trends are:

1) Increasing WER results in decreasing precision.
2) Longer queries are more robust to errors than shorter queries.

On each of tables 2-4, columns with increasing WER display a clear decrease in precision. This indicates a lack of redundancy among the query terms; with extra errors, critical search terms are lost (or misleading ones are inserted) and the precision decreases.

This trend holds for all three query lengths (2-4, 5-8, and 10-15 content words). But the relative decrease in precision was less for the longer queries. For 30.0% WER, the average loss for the "medium" length queries was 27.8%, while for the "short" queries it was 30.3%, and for the "very short" queries, 31.3%.

4. CONCLUSIONS

Examining the results in Experiment 2 in detail shows that most queries lost little precision, but that a few were damaged very badly. 5 of the 35 queries lost more than 50% average precision and two lost close to 100%. As a consequence, the median loss of precision 7.2% was much better than mean loss of 12.5%. The "trimmed mean" (used here as the average excluding the top and bottom 25 percentiles) was 6.69%. Though this indicates that in general performance may be even better than the mean figures would indicate, a manual examination of the queries that did badly shows that they will be hard to fix (i.e., there is nothing unusual about either the queries or the recognition errors in question.) However, because the damaging errors are apparently random, extending or rephrasing the query might well produce good results.

Experiment 3 demonstrates the same effect; that some queries suffer no decrease in precision (or actually increase), while others are disastrously degraded. If we consider the experiment with 30.0% WER (column 2 in Tables 2-4), and average over documents of all lengths, the mean precision decrease for 5 returned documents is 19.6%, while the "trimmed mean decrease" is only 13.6%.

Experiment 3 demonstrates that shorter queries are not as robust to errors introduced by the recognizer. For example, we find that introducing 30.0% word errors into 2-4 word queries degrades the precision at 10 documents by 33%, while the 10-15 word queries suffered a 28% decrease in precision. The stronger trend, however, is that increasing WER quickly decreases the precision. Experiment 1 indicated that long queries suffer only about a 10% drop in precision for a 28% WER; it is clear that by moving to shorter queries we have sacrificed a substantial amount of redundancy and therefore suffer larger losses of precision.

We are extending this work in several directions. We are investigating more effective way to combine the "n-best" list of hypotheses returned by the recognizer to improve retrieval precision. Experiment 3 simply used the concatenation of the top 5 hypotheses; it be may be more advantageous to include more hypotheses, or to weight the different hypotheses according to their acoustic or language scores. We are also developing a dialogue system capable of interacting with the user to help guide their search for documents (for example, by prompting for more search terms).

5. ACKNOWLEDGMENT

This work was supported by the NIST Advanced Technology Program, award 70NANB5H1181.

6. REFERENCES

[1] Callan J. P., Croft, W. B. and Broglio, J. *"TREC and TIPSTER Experiments with INQUERY."*, Information Processing and Management, (1994).

[2] Excalibur Technologies Corp. "Excalibur RetrievalWare", Columbia, Maryland, 1996.

[3] Harman, D. The DARPA TIPSTER Project. SIGIR Forum 26(2), 1992.

[4] Hauptmann, A. G. and Wactlar, H. D., *"Indexing and Search of Multimodal Information"*, Proceedings of ICASSP-97, pp. 195-198, Munich, 1997.

[5] James, D. *"A System for Unrestricted Topic Retrieval from Radio News Broadcasts"*, Proceedings of ICASSP-96, Atlanta, 1996.

[6] Jones, G. J. F., Foote, J. T., Sparck Jones, K. and Young, S. J. *"Video mail retrieval: The effects of word spotting accuracy on precision"*, Proceedings of ICASSP-95, Detroit, 1995.

[7] Jones, G. J. F., Foote, J. T., Sparck Jones, K. and Young, S. J. *"Robust Talker-Independent Audio Document Retrieval"*. Proceedings of ICASSP-96, Atlanta, 1996.

[8] Kupiec, J., Kimber, D., and Balasubramanian, V. *"Speech-based Retrieval Using Semantic Co-Ocurrence Filtering"*, Proceedings of the ARPA Human Language Technology Workshop 1994.

[9] Peskin, B., Connolly, S., Gillick, L., Lowe, S., McAllaster, D., Nagesha, V., van Mulbregt, P., and Wegmann, S. *"Improvements in Switchboard Recognition and Topic Identification"*. Proceedings of ICASSP-96, Atlanta, 1996.

[10] Wactlar, H. D., Hauptmann, A. G., and Witbrock, M. J., *"Informedia: News-on-Demand Experiments in Speech Recognition"*, Proceedings of the DARPA Speech Recognition Workshop, February 1996.

[11] Young, S. J., Brown, M. G., Foote, J. T., Jones, G. J. F., and Sparck Jones, K., *"Acoustic Indexing for Multimedia Retrieval and Browsing"*, Proceedings of ICASSP-97, pp. 199-202, 1997.