1998 TREC-7 Spoken Document Retrieval Track Overview and Results

John S. Garofolo, Ellen M. Voorhees, Cedric G. P. Auzanne, Vincent M. Stanford, Bruce A. Lund

National Institute of Standards and Technology (NIST)
Information Technology Laboratory
Building 225, Room A-216
Gaithersburg, MD 20899

ABSTRACT

This paper describes the 1998 TREC-7 Spoken Document Retrieval (SDR) Track which implemented an evaluation of retrieval of broadcast news excerpts using a combination of automatic speech recognition and information retrieval technologies. The motivations behind the SDR Track and background regarding its development and implementation are discussed. The SDR evaluation collection and topics are described and summaries and analyses of the results of the track are presented. Alternative metrics for automatic speech recognition as applicable to retrieval applications are also explored. Finally, plans for future SDR tracks are described.

1. BACKGROUND

Spoken Document Retrieval (SDR) involves the search and retrieval of excerpts from recordings of speech using a combination of automatic speech recognition and information retrieval techniques. In performing SDR, a speech recognition engine is applied to an audio input stream and generates a time-marked textual representation (transcription) of the speech. The transcription is then indexed and may be searched using an information retrieval engine. In traditional information retrieval, a topic (or query) results in a rank-ordered list of documents. In SDR, a topic results in a rank-ordered list of temporal pointers to potentially relevant excerpts. In an operational SDR system, these excerpts could be topical sections of a recording of a conference or radio or television broadcasts.

SDR was chosen as a TREC domain because of its potential use in navigating large multi-media collections of the near future and because it was believed that the component Automatic Speech Recognition and Information Retrieval technologies might work well enough now for usable SDR in some domains. SDR also provides a rich research domain in that it supports both development of large-scale near-real-time continuous speech recognition technologies and technologies for retrieval of spoken language. Further, SDR provides a venue for synergy between the speech recognition and information retrieval communities to improve both technologies and create hybrids.

The first community-wide evaluation SDR technology was implemented in 1997 for TREC-6. This pilot evaluation implemented a "known-item" task in which a particular relevant document was to be retrieved for each of a set of queries over a 50-hour collection of radio and television news broadcasts. Three retrieval conditions were implemented to examine the effect of recognition performance on retrieval performance:

1. *Reference* - retrieval using human-generated reference transcripts which for the purposes of this evaluation were considered to have "perfect" recognition.

2. *Baseline* - retrieval using IBM-contributed recognizer-generated transcripts with a 50% Word Error Rate. This provided both a common recognition error condition and an entrée for sites which did not have access to a recognition system of their own.

3. *Speech* - retrieval using the recordings of the broadcasts themselves requiring both recognition and retrieval technologies.

Thirteen sites participated in the pilot evaluation, eight of which implemented the Speech retrieval condition using their own or a team site's speech recognition system. The pilot evaluation proved that an evaluation of SDR technology could be implemented and that existing technologies worked quite well for a known-item task on a small collection. The results were so good that NIST chose to highlight the percent of target stories which were top-ranked (retrieved at rank one) by the systems.

Using the Percent Retrieved at Rank 1 metric, the University of Massachusetts retrieval system yielded the best performance for all three conditions. The UMass system achieved a retrieval rate of 78.7% for the Reference Retrieval condition and 63.8% for the Baseline Retrieval condition. For the Full SDR condition, UMass using a Dragon-Systems-produced 1-best recognizer transcript

with a 35% Word Error Rate, achieved a 76.6% retrieval rate.[1] The 2.1% difference in performance between retrieval using the reference transcripts and retrieval using the Dragon recognizer transcripts represented only one unretrieved story out of the 49 test topics.

2. MOTIVATION

The 1998 SDR Track was designed to address the known inadequacies in the 1997 SDR Track (small corpus, known-item task) to provide a more realistically challenging retrieval task. For 1998, an approximately 100-hour broadcast news test set collected by the Linguistic Data Consortium (LDC)[2], used previously as "the second 100 hours of BN training for Hub-4 recognition systems", was selected and a traditional TREC ad-hoc-style relevance task was chosen with topics and relevance assessments generated by human assessors. Two recognizer-produced transcript sets with different word error rates were provided by NIST as well as LDC human-generated reference transcripts. Also, for the first time, sites were encouraged to contribute their one-best recognizer-produced transcripts so that other sites could run retrieval on them. The improved test paradigm and alternative transcription sets with a spectrum of recognition error rates permitted us to further examine the relationship between recognition errors and retrieval accuracy. The new cross-recognizer task also permitted us to explore the development of alternative metrics for automatic speech recognition technology which would address particular inadequacies of the technology with regard to its use in information retrieval applications.

3. SDR EVALUATION PLAN

The complete evaluation plan for the 1998 TREC-7 Spoken Document Retrieval Track can be found at:

http://www.nist.gov/speech/sdr98/sdr98.htm

3.1 Evaluation Modes

The SDR Track included four retrieval conditions which provided component control experiments:

Reference (R1) (required) – Retrieval using the "perfect" human-transcribed reference transcripts of the Broadcast News recordings. This condition provided a control for retrieval.

Baseline (B1/B2) (required) – Retrieval using two sets of speech-recognition-generated 1-best transcripts produced by NIST using the CMU SPHINX-III recognition system. The Baseline-1 (B1) transcripts contained a moderate (33.8%) word error rate (relative to the current state-of-the-art) and the Baseline-2 (B2) transcripts contained a substantially higher (46.6%) word error rate. This condition provided two controls for recognition and permitted sites without access to recognition technology to participate.

Speech (S1/S2) (optional) – Retrieval using the Broadcast News recordings. This condition required both speech recognition and retrieval (which could be implemented by different sites). Two recognition/retrieval runs were permitted.

Cross Recognizer (CR) (optional) - Retrieval using 1-best speech-recognizer-generated transcripts contributed by other sites. This condition provided a control for recognition as well as allowing us to evaluate retrieval using a variety of recognition systems with a range of error rates.

One of the goals of the SDR Track is to encourage broad participation from both the Speech Recognition and Information Retrieval Communities. Therefore, the evaluation plan was designed to allow relatively easy entry for members of both communities. Speech recognition and retrieval experts were encouraged to team up to create pipelined or hybrid SDR systems. In addition, two participation levels were created to allow involvement by retrieval sites which did not have access to a speech recognition system:

Quasi-SDR - Sites without access to speech recognition technology were permitted to run retrieval on only the baseline recognizer transcripts and reference transcripts. (Retrieval conditions R1, B1, B2 minimally)

Full-SDR - Sites with access to speech recognition systems implemented both recognition and retrieval on the recorded news broadcasts as well as retrieval alone on the baseline recognizer transcripts and reference transcripts. (Retrieval conditions R1, B1, B2, and S1 minimally)

Participants in Full SDR with 1-best word-based recognizers were encouraged to submit their recognized transcripts to NIST. This provided the material to be used by other sites in implementing the Cross-Recognizer retrieval condition and permitted NIST to evaluate the effect of recognition error rates on retrieval performance.

For purposes of simplifying the implementation and evaluation process, the hand-annotated temporal story boundaries were given in all conditions. Figure 1. shows the general process for the TREC SDR task.

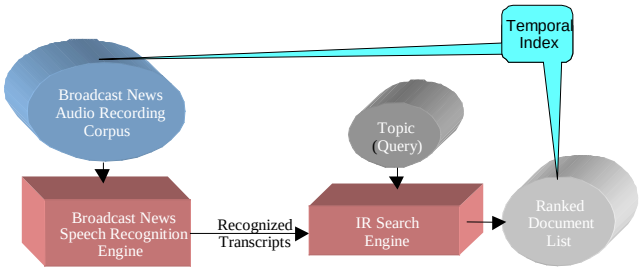

Figure 1. TREC SDR Process

3.2 Test Corpora

The LDC Broadcast News corpus was chosen for the SDR task since it contained news data from several radio and television sources and was fully transcribed and pre-segmented by story.[2] To adapt the BN corpus to the SDR task, story ID tags were added to uniquely identify each annotated story for retrieval and scoring.

A subset of 100 hours of the Broadcast News Corpus collected between June 1997 and January 1998 (which was originally collected by the LDC to provide training material for DARPA Hub-4 speech recognition systems) was chosen as the test corpus. The corpus was filtered to exclude commercials, sports summaries, weather reports, and untranscribed stories. In all, 87 hours of the 100-hour subset were selected as the test collection for the SDR evaluation. Because the story boundaries were to be known, an index giving the story IDs and time of each story boundary was provided to test participants.

The final filtered test set contained 2,866 stories with about 772,000 words. Roughly 1/3 of the stories in the test set were labeled as "filler" – non-topical sections of the broadcasts. Because of the small size of the collection for retrieval testing, these were not removed from the test set. The mean length in words for the stories in the test set was 269 words. The histogram in Figure 2 shows the distribution of the length of the stories in the test set. Note that about half of the stories contain less than 100 words and a few stories contain 2000 or more words.

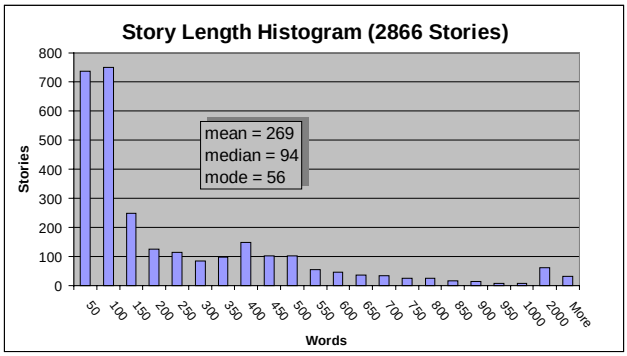

Figure 2. Test Collection Story Length Histogram

The recorded waveform material for the Speech retrieval condition was made available to the participants in April, 1998. The human-created reference transcripts for the test collection and indices which specified the 2,866 usable stories were released in June. The test topics and baseline recognizer transcripts were released in the beginning of July and results were due at NIST at the end of August. The results of the SDR track were reported at TREC-7 in November 1998 and at the DARPA Broadcast News Workshop in March 1999.

3.3 Baseline Recognizer Transcripts

CMU permitted NIST to use its SPHINX-III broadcast news recognition system to create a set of recognition-generated transcripts for the baseline retrieval (B1) condition. Since SPHINX-III ran in nearly 200 times real time on NIST's UNIX-based workstations, NIST realized that it would take almost two years of computation to complete one recognition pass over the 87 hours of recordings in the SDR test collection. NIST learned of inexpensive clusters of PC-based LINUX systems being used by NASA in its BEOWULF [3] project and set out to create such a system for recognition so that it could parallelize the recognition task.

An architecture was created in which a single scheduling server was used to control 8 computation nodes each containing 200-MHz Pentium Pro processors with 256 Mb. of memory and a 1 Gb. disk drive for swapping. The nodes were set up to boot from the server and both the server and nodes used the LINUX 2.0.32 operating system.

To implement distributed recognition, a CMU-contributed segmenter was first run on the recordings to break them up into tractable chunks of about 45 seconds each for recognition. A network scheduler using a FIFO-with-priorities algorithm (GNQS) was used to queue and track the chunks for processing over the available nodes.[4] The scheduling and network overhead was relatively low, so

the cluster performed roughly 8 times faster than a single processor machine.

The NIST High Performance Systems and Services (HPSS) Division was also investigating the use of such clusters as an alternative to supercomputers in servicing the computational needs of the NIST measurement laboratories and allowed us to enlist their nodes. This gave us access to 32 additional nodes and permitted HPSS to measure the performance of the technology. In all, 40 nodes were employed to create the B1 transcripts.

With 40 nodes, NIST was able to implement 2 baseline recognition runs. The first (B1) run was implemented with SPHINX running at moderate accuracy, using only the forward Viterbi search. This system benchmarked with the NIST SCLITE scoring software at 27.1% word error rate on the Hub-4 '97 test set and at 33.8% word error rate on the SDR test collection. NIST decided to create a second, less optimal run to examine the effect of recognition degradation on retrieval performance. The second (B2) run implemented the same SPHINX system, but with its pruning thresholds lowered. This system benchmarked at 46.6% word error rate on the SDR '98 test collection (comparable to the 50% word error rate for the IBM recognition system used to create the baseline recognizer transcripts in the 1997 SDR evaluation. [1])

3.4 SDR Topics

A team of 3 NIST TREC assessors met in April 1998 to select 25 topics for the test collection using similar procedures to those used in other TREC ad-hoc tasks. The assessors were instructed to find topics with 7 or more relevant news stories each in the collection using the NIST PRISE search engine. Unlike 1997, the assessors were not instructed to artificially construct the topics to exercise a particular component of the SDR systems. Because of content limitations in the collection, however, the assessors were able to develop only 23 usable topics including:

> *Find reports of fatal air crashes* (Topic 62)
>
> *What economic developments have occurred in Hong Kong since its incorporation into the Chinese People's Republic* (Topic 63)

As in other TREC evaluations, once the retrieval results were submitted to NIST, the output of the participating systems was used to create pools of stories to be evaluated for relevancy by the assessors. The pools were created by taking the union of the top 100 stories for each topic output by each of the systems for each of the R1, B1, B1, B2, S1, and S2 retrieval conditions. The assessors met again in September and exhaustively examined the pools to create a reference set of relevant documents for each topic which was then used to score the results of the evaluation. Figure 3 shows the relevancy profile for the test collection with regard to the test topics.

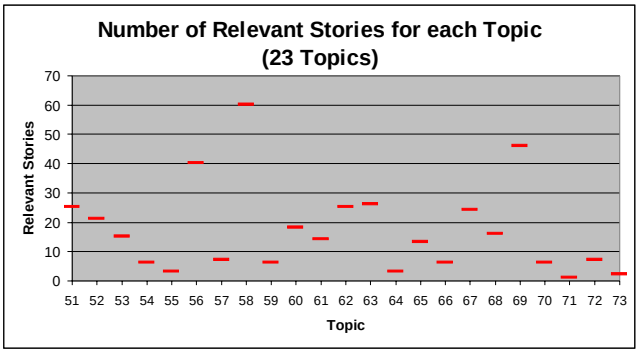

Figure 3. Relevant Stories Per Topic

On average, there were 17 relevant stories per topic. However, as Figure 3 shows, there was a great deal of variability in the number of relevant stories for particular topics. For instance, at the extremes, Topic 71 had only 1 relevant story and Topic 58 had 60 relevant stories.

4. EVALUATION RESULTS

In all, 11 sites (or recognition/retrieval teams) participated in the SDR Track. Eight of these sites performed the Full SDR task by implementing both the recognition and retrieval components of the task (S1). These sites were also required to implement the R1, B1, and B2 control conditions.

Full SDR (recognition and retrieval - R1, B1, B2, S1):

- AT&T (ATT)
- CMU Group 1 (CMU1)
- Cambridge University, UK (CUHTK)
- DERA, UK (DERA)
- Royal Melbourne Institute of Technology, Australia (MDS)
- Sheffield University, UK (SHEF)
- TNO-TPD TU-Delft, Netherlands (TNO)
- University of MA - Dragon Systems (UMass)

AT&T, CMU Group 1, DERA, RMIT, and UMass implemented secondary (S2) recognition/retrieval systems, although only DERA submitted their secondary recognition system output for scoring and redistribution.

Four of the Full SDR sites (Cambridge, DERA, RMIT, and Sheffield also implemented the Cross-Recognizer (CR) condition.

The remaining 3 sites performed only the Quasi-SDR portion of the task.

Quasi-SDR (retrieval only - R1, B1, B2):

- CMU Group 2 (CMU2)
- NSA (NSA)
- University of MD (UMD)

4.1 Speech Recognition Component Performance

The primary purpose of the SDR Track was to evaluate the retrieval of spoken documents. To this end, there was not a formal evaluation of the speech recognition component of the Full SDR systems. However, if sites used 1-best word recognition to produce transcripts as input to their retrieval systems, they were encouraged to submit these for sharing in the Cross-Recognizer condition and for NIST evaluation of the effect of recognition performance on retrieval.

It should be noted that the SDR recognition error rates are not directly comparable to error rates obtained in the NIST Hub-4 Broadcast News Transcription tasks, since the intensive verification and orthographic normalization performed for Hub-4 transcripts are not performed for SDR transcripts. Because of this, the word error rates obtained for SDR will be somewhat higher than for the identical system run on a Hub-4 test set. As a case in point, the CMU SPHINX-III-based recognition system implemented at NIST to create the transcripts for the Baseline 1 (B1) retrieval condition was scored with a 33.8% word error rate against the SDR reference transcripts. However, the identical system was benchmarked at 27.7% word error rate using the 1997 Hub-4 test set. [5]

Of the 8 participating Full SDR sites, 5 submitted recognition output to NIST for scoring. Other Full SDR sites either used an alternative recognition technique such as phone-based recognition or word lattices or chose not to share their recognition results. Figure 4 shows a frequency plot in profile of the story word error rates for each of the submitted 1-best systems. This plot gives a graphical profile of recognizer performance over the 2866 stories in the collection.

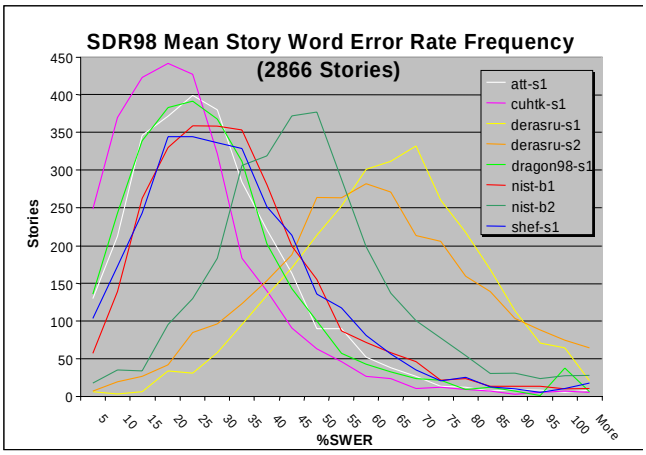

Figure 4. Story Word Error Rate Frequency Plot in Profile for Submitted Recognized Transcripts

Figure 5 shows the mean story word error rate (SWER) and mean test set word error rate (WER) for each of the submitted recognition systems. The ovals indicate no significant difference between systems in mean story word error rate error rate at 95% confidence. The best recognition results were from the Cambridge University HTK recognition system with a 24.6% test set word error rate and a 22.2% mean story word error rate.[6] A complete table of recognition scores for the submitted systems is given in Appendix A. Note that the results are slightly improved over what was reported at the TREC-7 meeting. After the meeting, it was found that there were severe story boundary annotation errors in 5 of the stories which yielded extremely high word error rates for those stories (on the order of 2,000%). These annotations were corrected and all of the results rescored.

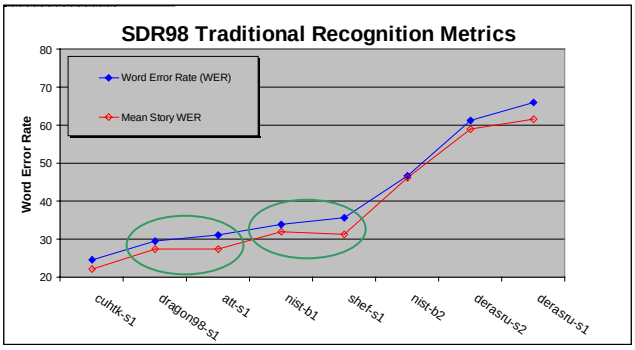

Figure 5 . Test Set and Mean Story Word Error Rate for Submitted Recognized Transcripts with Cross-System Significance at 95% for SWER

The submitted recognition systems exhibited a wide range of error rates and provided a spectrum of material for the Cross-Recognizer retrieval experiment.

4.2 Retrieval Results

Test participants were required to submit a relevance-rank-ordered list of the ID's of the top 1000 stories they retrieved for each topic. These results were then scored against the reference assessments created by the NIST assessors using the TREC_EVAL scoring software. As in other TREC tasks, the primary retrieval metric for the SDR evaluation was mean average precision over all topics. Mean average precision (MAP)1 is the metric employed in TREC retrieval tracks to provide a single figure of merit.[7] Figure 6 shows the MAP results for participating retrieval systems for the R1, B1, B2, S1, and S2 retrieval conditions.

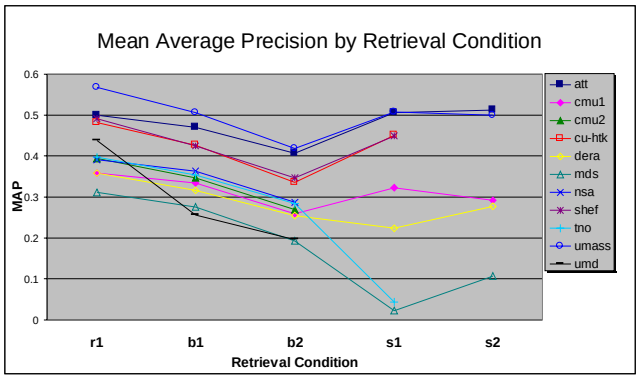

Figure 6 . Mean Average Precision for Required Retrieval Conditions

The graph shows that for all retrieval conditions except S2, the University of Massachusetts system achieved the best mean average precision. The UMass system achieved a MAP for retrieval using the human reference transcripts (R1) of .5668. The same system's retrieval for the moderate error baseline recognizer (B1) transcripts achieved a MAP of .5063, and for the high error baseline recognizer (B2) transcripts, a MAP of .4191. For the Speech input condition (S1) using their own team site's (Dragon Systems) recognizer at 29.5% word error rate, the UMass system achieved a MAP of .5075.[8] The AT&T system performed similarly for the S1 condition with a MAP of .5065. AT&T implemented a second recognition/retrieval system (S2) which achieved a MAP of .5120 - the highest results for input from a recognizer in this evaluation. It is interesting to note that the AT&T S1 and S2 results exceeded the results AT&T obtained (.4992 MAP) for the human reference transcripts (R1). AT&T attributes this to a new approach they implemented for document expansion using contemporaneous newswire texts. They applied the new document expansion approach only to their S1 and S2 runs and not to their other runs.[9] Appendix A gives a complete tabulation of the mean average precision scores for all of the systems and conditions.

In general, the results for this evaluation were quite good, with a near-linear decline in mean average precision for recognition transcripts with higher word error rates. The Cross-Recognizer retrieval results were used to further explore this apparent relationship.

4.3 Cross-Recognizer Retrieval Results and Alternative Recognition Metrics

This year, sites were encouraged to share their recognizer transcripts with other retrieval sites to implement a cross-recognizer retrieval condition. This cross testing permitted the examination of retrieval performance over a wider variety of recognized transcripts and we could begin to truly examine the relationship between recognition performance and retrieval performance. It also provided us with data to evaluate our recognition metrics for their suitability for retrieval and to experiment with new ones as well.

Four of the Full SDR sites: Cambridge University, DERA, RMIT/MDS, and Sheffield University implemented the cross-recognizer (CR) retrieval condition. The CR condition provided 9 recognition/retrieval points: the reference transcripts with "perfect" recognition, the baseline B1 and B2 transcripts, and 6 other recognizer-generated transcripts contributed by 5 sites: AT&T, Cambridge-HTK, DERA (2 sets), Dragon Systems, and Sheffield University. These recognized transcripts covered a wide range of word error rates.

When we plot mean average precision against mean story word error rate for each of the 4 retrieval systems (Figure 7), we see a linear trend in mean average precision as average recognition word error increases. The correlation averaged over these 4 retrieval systems for the 9 recognition points is ≈ .87. This high correlation indicates that there is indeed a significant relationship between word error rate and retrieval accuracy. The plot also shows a consistent pattern in performance profiles across the retrieval systems with respect to recognizers. However, the retrieval results for all systems for the B2 recognizer are low with respect to the word error rate metric (much lower than the results for the two DERA recognizers with significantly higher word error rates.) This tells us that word error rate alone is insufficient to fully predict retrieval performance.

1 Mean Average Precision (MAP) is a composite measure of retrieval performance and is equivalent to the mean across topics of the area under the uninterpolated precision/recall graph for each topic.

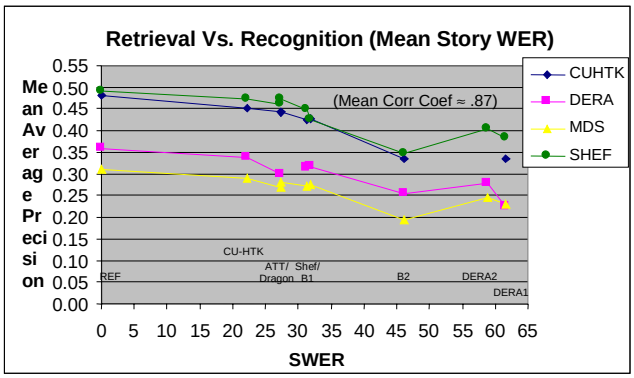

Figure 7. Cross-Recognizer Results: Mean Average Precision vs. Mean Story Word Error Rate

We believed that we might achieve an even higher correlation if a metric for speech recognition were employed which emphasizes the information-carrying words which are key for retrieval. Such a metric would be more predictive of retrieval performance and could be used to determine the suitability of a recognizer for use in a retrieval task.

We considered 3 types of metrics:

Named-Entity-based: This metric evaluates the error rate for named-entity words (people, locations, and organizations) as defined in the 1998 Hub-4 Information Extraction - Named Entity (IE-NE) Evaluation.[10] The disadvantage of this metric is that it requires named-entity annotations in the reference transcripts. Fortunately, GTE/BBN had annotated the SDR reference transcriptions for use as named-entity training data for the IE-NE evaluation.[11] We developed the following metric:

> *named entity word error rate* (ne-wer): score only the named entities in the recognizer transcripts. To implement ne-wer, we used IE-Eval/REEP Named Entity scoring software [12] to align the annotated named entity words in the reference transcript with words in the recognizer transcripts. The alignments (with embedded named-entity tags) were then scored using the NIST SCLITE speech recognition scoring software. The embedded tags permitted us to score only named-entity words. So as not to introduce entity tagger error into our metric, we ignored named entity words which might be inserted by the recognizer and evaluated only named entity words as annotated in the reference transcripts.

General IR-based: These metrics use IR approaches themselves to process, filter, and weight the words in the recognizer transcripts to be scored. Such metrics are potentially useful in predicting retrieval performance based on recognition performance and might, therefore, be used to tune a recognizer for a retrieval task. We considered 3 such metrics:

> *stop-word-filtered word error rate* (swf-wer): apply a stop-word list to the words in the reference and recognizer transcripts to remove stop (non-information-carrying) words. To implement this metric, we removed all occurrences of words in a 396-word stop word list from both the reference and recognizer transcripts. We then performed SCLITE word error rate scoring on the filtered transcripts.
>
> *Stemmed stop-word-filtered word error rate* (sswf-wer): apply a stemmer to the results of the swf-wer filtering process above to remove word differences which are irrelevant to retrieval algorithms. To implement this metric, we applied an implementation of the Porter stemmer [13] to the stop-word-filtered reference and recognizer transcripts. We then performed SCLITE word error rate scoring on the filtered transcripts.
>
> *IR-weighted stemmed stop-word-filtered word error rate* (IRW-WER). Apply an IR indexing algorithm to weight words prior to SCLITE word error rate scoring. We are currently examining IR algorithms for this application and have not yet implemented this metric.

Query-Set-Specific: These metrics evaluate the word error rate only on words given in the test topics. Such metrics are useful in analyzing the results of a given test, but are not predictive. We considered the following metric:

> *query-word word error rate* (QW-WER). To implement this, we identify the words in the test topics, remove stop words, and stem the remainder. The reference transcripts are also stemmed. The processed query word list is then used to score only the occurrences of these words in the processed reference transcripts against the corresponding aligned words in the processed recognizer transcripts. This metric has not yet been implemented.

The results of these alternative metrics as applied to the SDR recognizer are shown in Figure 8. Note that only the Named-Entity-based metrics clearly change the relative

ranking of the recognizer transcript sets. These two metrics show that the B2 recognition system was a poorer performer with regard to named entities than is evidenced by its word error rate. Our hypothesis is that the adjustment we made to the SPHINX pruning thresholds artificially reduced the likelihood of longer words being recognized - words which are more likely to be content-carrying named entities. Surprisingly, the other recognition metrics don't seem to be significantly different than word error rate in measuring recognition performance - an indication that recognition systems perform just as well (or poorly) on content words than on non-content words. This contradicts popular folklore that speech recognition systems perform more poorly on non-content-bearing "function" words. The scores for each of the metrics as applied to each of the recognized transcripts are given in Appendix A.

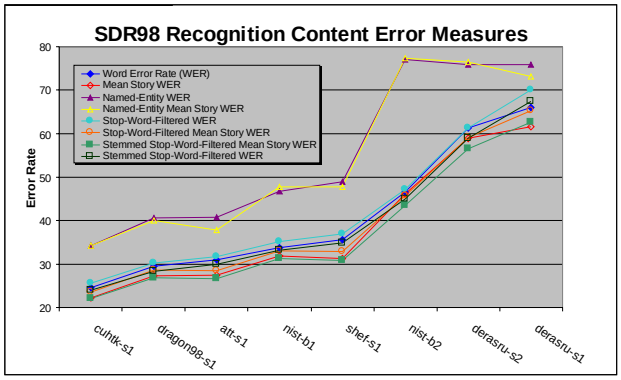

Figure 8 . Alternative Recognition Metrics

To quantify the efficacy of these metrics as predictive tools, we display a correlation analysis of the scores for the 4 retrieval systems versus the recognition metrics for each of the 9 transcript sets in Table 1.

	CUHTK	DERA	MDS	SHEF	Mean Corr	Mean Rank
WER	-0.901	-0.905	-0.785	-0.797	-0.847	6.00
SWER	-0.927	-0.912	-0.812	-0.827	-0.869	3.25
NE-WER	-0.937	-0.900	-0.897	-0.890	-0.906	3.00
NE-SWER	-0.936	-0.886	-0.900	-0.898	-0.905	3.00
SWF-WER	-0.894	-0.911	-0.777	-0.791	-0.843	7.00
SWF-SWER	-0.911	-0.915	-0.794	-0.811	-0.858	4.00
SSWF-WER	-0.897	-0.913	-0.776	-0.793	-0.845	6.25
SSWF-SWER	-0.914	-0.916	-0.794	-0.812	-0.859	3.25

Table 1 . Correlation Between Recognition Metrics and Retrieval Performance

The table shows that, on average for all 4 retrieval systems, the named entity test set word error rate (ne-wer) and named entity mean story word error rate (ne-swer) metrics provide the best correlation with retrieval performance with mean system correlation coefficient values of .906 and .905 and with minimal (best) mean ranks of 3.0 derived from the individual system correlation coefficients. The CU-HTK, MDS, and Sheffield retrieval systems are all most highly correlated with the named-entity-based metrics. However, the DERA retrieval system seems to be a bit of an outlier since it is more correlated with stemmed stop-word-filtered mean story word error rate (sswf-swer). In any case, all of the metrics including the traditional word error rate metrics are significantly correlated with retrieval performance.

The high correlation between named entity mean story word error rate (ne-swer) and retrieval performance is visually depicted in Figure 9.

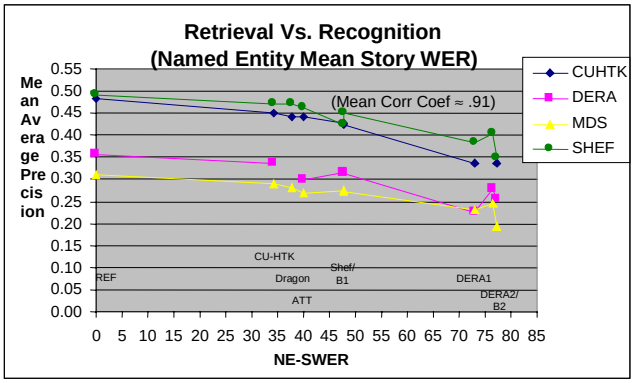

Figure 9. Cross-Recognizer Results: Mean Average Precision vs. Named Entity Mean Story Word Error Rate

We have yet to perform the scoring using the IR-weighted metric. We believe that it may provide a better predictor of retrieval performance than simple word error rate without the cost of annotation associated with named entity word error rate.

5. CONCLUSIONS

In 1997, we found that we could successfully implement a known-item retrieval task using broadcast news. In 1998, we found that we could successfully implement an ad-hoc retrieval task using a larger corpus of broadcast news. The best performance for retrieval using a speech recognizer (.5120 MAP) approached the best performance for retrieval using perfect human-generated reference transcripts (.5668).

We also found that there is a near-linear relationship between recognition word error rate and retrieval performance. We investigated alternative metrics for recognition performance that might be more predictive of retrieval performance. We found that named-entity word error rate was more highly correlated with retrieval performance than word error rate alone. We intend to continue investigating IR-based algorithms that could be applied to building recognition metrics tuned for retrieval

applications.

However, we are hesitant to declare retrieval using speech recognition-generated transcripts a solved problem. The 1998 SDR collection of 2,866 stories is still quite small for retrieval evaluation. The next challenge is to determine how well retrieval performance scales for larger realistic collections of broadcast news and to remove artificially constrained components of the evaluation such as known story boundaries.

6. FUTURE

The 1998 SDR track employed a corpus which was twice as large as that used for the 1997 track. However, the corpus was not collected for retrieval purposes and is not appropriately representative with regard to sources and time. For 1999, we plan to use a 5-month subset of the TDT-2 corpus for the SDR evaluation. The TDT-2 corpus, which was collected by the Linguistic Data Consortium for the DARPA Topic Detection and Tracking Tasks, contains 632 hours/24,503 stories of broadcast news from ABC World News Tonight, CNN Headline News, PRI The World, and several Voice of America programs. It is well-suited for the SDR task in that the broadcast news sources are evenly sampled over a 6-month time period from January through June, 1998.[14] Further, it contains a complementary newswire corpus which can be used by sites who wish to explore the application of rolling language models to the SDR recognition task. Rolling language models evolve over time to changes in language and will address real-world use of recognition for time-continuous tasks. Next year's evaluation will support two language-modeling modes: fixed and rolling.

The TDT-2 corpus does not have Hub-4-style transcripts, but does have closed-caption transcriptions for the television programs and comparable quality transcripts for the radio programs. A minimum of 10 hours of randomly selected stories in the corpus will be transcribed in the Hub-4 style so that speech recognition performance can be benchmarked. It is hoped, however, that the entire corpus will eventually be transcribed so that more extensive benchmarking can be performed.

The NIST assessors will create 50 ad-hoc-style topics for the 1999 SDR track using the existing transcripts. These transcripts will also be used in the reference condition for the evaluation.

There was increased interest at TREC-7 in supporting an evaluation condition in which story boundaries are unknown, which would more naturally model a real implementation of SDR. To support this condition, systems will be permitted to make an optional run on the baseline speech recognizer transcripts and their own recognizer output without story boundaries. The systems will output a time stamp for the top 1000 retrieved stories rather than a story ID. Each time stamp will be mapped to the story that contains it. Duplicate stories will be eliminated, so that systems which output multiple time stamps referring to the same stories will be penalized. The inferred story IDs will then be used to implement traditional TREC_EVAL scoring. In this task, systems should attempt to find the mid-point or "hotspots" in stories. This approach is simpler than other possible schemes that might require systems to output story boundaries and which are scored on distances. However, since story segmentation is not a focus of this track, and since it is desirable to have comparable results for both story-boundary-known and story-boundary-unknown conditions, this has been determined to be the most expeditious approach.

7. ACKNOWLEDGEMENTS

The authors would like to thank Karen Sparck Jones of Cambridge University for her tireless efforts in promoting the SDR Track and for her assistance in developing the evaluation plan. We'd also like to especially thank CMU for their contribution of the SPHINX-III recognizer for use as SDR the baseline recognizer. CMU's Kristie Seymore, Mosur "Ravi" Ravishankar, and Matt Siegler were of great help in getting SPHINX up and running at NIST. Finally, We'd like to thank Gilbert Sanseau of the NIST NLPIR Group for his suggestions of IR algorithms in our exploration of alternative speech recognition metrics.

NOTICE

REFERENCES

[1] Voorhees, E., Garofolo, J., Stanford, V., and Sparck Jones, K., *TREC-6 1997 Spoken Document Retrieval Track Overview and Results*, Proc. TREC-6, 1997 and 1998 DARPA Speech Recognition Workshop, February 1998.

[2] Graff, D., Wu, Z., MacIntyre, R., and Liberman, M., The 1996 Broadcast News Speech and Language-Model Corpus, Proc. DARPA Speech Recognition Workshop, February 1997.

[3] BEOWULF Project, NASA Center of Excellence in Space Data and Information Sciences, http://cesdis.gsfc.nasa.gov/linux/beowulf/

[4] *The Generic NQS Web Site - OpenSource Batch Processing*, http://www.gnqs.org/

[5] Pallett, D.S., Fiscus, J.G., Martin, A., Przybocki, M.A., *1997 Broadcast News Benchmark Test Results: English and Non-English*, Proc. DARPA Broadcast News Transcription and Understanding Workshop, February 1998.

[6] Johnson, S.E., Jourlin, P., Moore, G.L., Sparck Jones, K., Woodland, P.C., *Spoken Document Retrieval for TREC-7*, Proc. TREC-7, November 1998.

[7] Voorhees, E.M., Harman, D., *Overview of the Seventh Text REtrieval Conference (TREC-7)*, Proc. TREC-7, November 1998.

[8] Allan, J, Callan, J., Sanderson, Xu, J., *INQUERY and TREC-7*, Proc. TREC-7, November 1998.

[9] Singhal, A., Choi, J., Hindle, D., Lewis, D.D., Pereira, F., *AT&T at TREC7*, Proc. TREC-7, November 1998.

[10] Przybocki, M.A., Fiscus, J.G., Garofolo, J.S., Pallett, D.S., *1998 Hub-4 Information Extraction Evaluation*, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

[11] Miller, D., Schwartz, R., Weischedel, R., Stone, R., Named Entity Extraction from Broadcast News, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

[12] Douthout, A., *Hub-4 1998 IE-NE Scoring Software with Recognition and Extraction Evaluation Pipeline*, SAIC, ftp://jaguar.ncsl.nist.gov/csr98/official-IE-98_scoring.tar.Z

[13] Porter, M.F., An algorithm for suffix stripping, Program 14 (3), July 1980, pp. 130-137.

[14] Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S, TDT-2 Text and Speech Corpus, Proc. 1999 DARPA Broadcast News Workshop, March 1999.

Appendix A: 1998 TREC-7 Spoken Document Retrieval Track Summary Results

Retrieval Results - Mean Average Precision

Site	R1	B1	B2	S1	S2	CR-ATT	CR-CUHTK	CR-DERA1	CR-DERA2	CR-Dragon	CR-Shef
ATT	0.4992	0.4700	0.4065	0.5065	0.5120	0.5065					
CMU1	0.3577	0.3345	0.2590	0.3224	0.2926						
CMU2	0.3936	0.3472	0.2693								
CUHTK	0.4817	0.4272	0.3352	0.4509		0.4419	0.4509	0.3352		0.4428	0.4251
DERA	0.3579	0.3164	0.2551	0.2242	0.2768		0.3375	0.2242	0.2768	0.2990	0.3134
RMIT-MDS	0.3107	0.2753	0.1937	0.0223	0.1063	0.2812	0.2906	0.2309	0.2443	0.2704	0.2730
NSA	0.3907	0.3640	0.2868								
SHEF	0.4916	0.4243	0.3471	0.4495		0.4717	0.4713	0.3836	0.4047	0.4613	0.4495
TNO	0.3970	0.3533	0.2833	0.0436							
UMass	0.5668	0.5063	0.4191	0.5075	0.5000						
UMD	0.4386	0.2557	0.1967								

Speech Recognition Results - Various Metrics (%error)

ASR Metric	R1	B1	B2			CR-ATT	CR-CUHTK	CR-DERA1	CR-DERA2	CR-Dragon	CR-Shef
WER	0.0	33.8	46.6			31.0	24.6	66.0	61.3	29.5	35.6
SWER	0.0	31.9	46.1			27.4	22.2	61.6	58.9	27.3	31.3
NE-WER	0.0	46.8	77.1			40.8	34.2	75.9	75.9	40.6	49.0
NE-SWER	0.0	47.7	77.3			37.8	34.2	73.1	76.5	40.1	47.7
SWF-WER	0.0	35.1	47.2			31.8	25.7	70.0	61.3	30.2	37.0
SWF-SWER	0.0	33.1	45.7			28.4	23.6	65.3	59.0	28.6	32.9
SSWF-WER	0.0	33.2	44.9			29.9	24.0	67.4	58.9	28.3	34.8
SSWF-SWER	0.0	31.3	43.5			26.7	22.1	62.6	56.6	26.8	30.8