EURASIP Journal on Applied Signal Processing 2003:2, 103–114

Robust Techniques for Organizing and Retrieving Spoken Documents

James Allan

Center for Information Retrieval, Department of Computer Science, University of Massachusetts, Amherst, MA 01003-9264, USA
Email: allan@cs.umass.edu

Received 5 April 2002 and in revised form 7 November 2002

Information retrieval tasks such as document retrieval and topic detection and tracking (TDT) show little degradation when applied to speech recognizer output. We claim that the robustness of the process is because of inherent redundancy in the problem: not only are words repeated, but semantically related words also provide support. We show how document and query expansion can enhance that redundancy and make document retrieval robust to speech recognition errors. We show that the same effect is true for TDT's tracking task, but that recognizer errors are more of an issue for new event and story link detection.

Keywords and phrases: spoken document retrieval, topic detection and tracking, information retrieval.

1. INTRODUCTION

The prevalence and success of search engines on the Web have broadly illustrated that information retrieval (IR) methods can successfully find documents relevant to many queries. Given a brief description of what interests a searcher, retrieval services on and off the Web are generally able to provide a list of possibly matching items with acceptable accuracy. To be sure, the systems make plenty of mistakes and there is substantial room for improvement. However, retrieval of text is adequate for many types of searching.

As disk space has dropped in price and network bandwidth has improved, it is easier for people to move beyond text, using audio, images, or video as a key means for communicating information. With appropriate compression, it is also reasonable to expect that people and organizations will store large numbers of multimedia documents on their computer and will want to manage them in the same way written documents are managed.

"Managing" those documents includes the ability to search them to find something of interest to the same extent that text documents can be managed by a search engine. Although some of those multimedia documents will have metadata associated with them, and that metadata can be the foundation for some types of search, multimedia documents that include speech have words in them and in theory can be treated identically to text. That is, if the speech is converted to text, then all of the text indexing and retrieval techniques should in theory carry over to this class of multimedia document. In this paper, we discuss the impact of speech recognition systems on document organization and retrieval.

Text is, of course, present in some images also, either because it is a scanned document that can be character recognized [1] or because the picture chances to include some text in a sign or something like that [2]. Such text could also be extracted, and we expect that many of the observations that follow would carry over. However, we focus on speech documents in the discussion below.

One problem with the approach of converting the speech to text is that automatic speech recognition (transcription) systems are not perfect. They occasionally generate words that sound similar to what was said, sometimes drop words, and occasionally insert words that were not there.

Nevertheless, we claim that existing research in the field of information retrieval suggests that any problem related to that can be addressed with simple techniques. As a result, the ability of a system to provide accurate search results is not substantially affected by speech recognition errors.

We support this claim by exploring two information retrieval tasks and the impact of automatic speech recognition (ASR) errors on their accuracy. In Section 4, we discuss the problem of spoken document retrieval (SDR), finding speech documents in response to a query. In Section 5, we consider the problem of organizing broadcast news stories (audio) by the events that they describe, a set of tasks within topic detection and tracking (TDT). Before doing that, in Section 2, we will motivate the question by demonstrating that ASR errors look like they should be causing a problem. Then, in Section 3 we discuss techniques that are used in both SDR and TDT to compensate for ASR errors. We provide some counterpoint to the success of these techniques in Section 6 where we suggest problems for

Human transcription. In Massachusetts a common **bacteria** has turned into a killer. Children's Hospital in Boston says a new and aggressive strain of **pseudomonas** is to blame for the deaths of four newborns this summer in its neo-natal intensive care unit. All the babies had been critically ill before they were infected. That neo-natal unit has been closed since last month, but is it is expected to reopen sometime this week. Researchers continue to investigate the **bacteria** outbreak there.

30% errors. in massachusetts a common **bacteria** has turned into a killer. children's hospital in boston says the new and aggressive strain of **sue lawless** is to blame for the deaths of four newborns this summer in its neonatal intensive care you. all the babies had been critically ill before they were infected. and neonatal unit has been closed since last month but is it is expected to reopen sometime this week. researchers contain investigate the **bacteria** outbreak their.

45% errors. in massachusetts the common **bacteria** has turned into a killer. children's hospital in boston says a new one and aggressive strain of **somalis** used to blame for the deaths of four new boards this summer and steel nato intent security that. all the babies had been critically ill before they were infected. did you build unity has been closed since last month but it is expected to reopen sometime this week. researchers continue to investigate the **bacteria** up with their.

65% errors. it in massachusetts and common **bank syria** has turned in school killer. children's hospital and boston says the new and west slate of **civil list** is only the deaths for new ones this summer women's neo middle instances here you life. all the babies had been critically killed before they were effective. in new unit has been close since last month with israel is expected to wield some time this week. researchers continue to investigate the **bacteria** with enough.

Figure 1: A small extract from the beginning of TREC-7 SDR document em970915.4. The top document is from a human transcript and has been edited here to include punctuation for readability. The remaining three passages are the output of ASR systems on the corresponding audio, converted to lower case and with punctuation manually inserted to make them easier to read. They correspond to word error rates of approximately 30%, 45%, and 65%, in that order [3].

certain text processing capabilities. We present our conclusions in Section 7.

2. WHY ASR LOOKS LIKE A PROBLEM

The basis of most information retrieval technologies is word matching. Substantial effort goes into deciding how to model the problem, which words to include, how to weight them, how to introduce synonyms or other related words, and so on. Ultimately, however, at the core of most technologies is the idea that documents containing query words are more likely to be relevant to the query. It is this central concept that suggested speech recognition output might require complex processing.

ASR systems convert audio into words using a process that is far too involved to detail here [4]. The important point is that ASR systems make errors for a variety of reasons ranging from loud background noise, poor quality microphones, slurred speech, and words that are not in the system's vocabulary. When an ASR system makes a mistake, it drops words, inserts words, or selects a different set of words to include. If those words are critical for a search or other task, the error could be catastrophic.

To illustrate the problem, consider the passages of text in Figure 1. Those passages were taken from the TREC-7 spoken document retrieval test documents [3] discussing infant deaths from a deadly strain of a bacterium. Imagine a query that was looking for articles discussing *pseudomonas*: none of the ASR systems recognized that word (it may have been out of every system's vocabulary) and replaced it with their best guess, including the name of people from Somalia! Even the simpler word *bacteria* is confused by systems with higher word error rate.

The best ASR systems are currently achieving about 10% word error rate for speakers reading from a script into studio-quality microphones. When the background is noisy or the speech is conversational, the error rates are closer to 30–40%, yielding errors such as those in Figure 1. This suggests that for most speech to be retrievable, IR technologies need to be enhanced to cope with errors introduced by speech recognition systems.

3. ADDRESSING ASR ISSUES

The basic approach that IR and organization systems use to minimize problems from ASR errors is to expand the query or the documents. That is, because the problem is that critical words are missing, the goal of the techniques is to bring the words back through some other means.

There are two basic approaches to this. The first is to leverage information from the speech recognition system itself and to include various words that it thought were possible, but not the most probable choices. The second is to use corpus-based statistical approaches to expand the query or the documents with words that are related to the starting text. Both techniques work, though using the information from

Human transcription
the point of **state** aid is **fairness** that is to provide an **adequate education** to each **child** not **withstanding** the **income** of the child's family

Top three hypotheses
hate fair adequate education child withstand calm
hate fair adequate education child withstand common
hate fair adequate education child withstand intercom

Figure 2: Text from TREC-6 test document j960531d.7 along with three likely decodings (after being stopped and stemmed) proposed by one recognizer [5]. Note that the word "income" has the most variability here and that "state" was consistently misrecognized.

the recognizer seems more error prone and is less portable because it requires additional output from the speech recognition system.

3.1. Recognizer-based expansion

Broadly speaking, speech recognition systems work by generating an inter-connected lattice of possible words that might have been uttered. The nodes in the lattice are words with probabilities associated with them, and they are connected in the order they might have been spoken. One of the last parts of the ASR process is to scan through the lattice and select the best possible (i.e., most probable) sequence of words from the lattice of hypotheses.

We know that ASR systems make mistakes and choose words that, although the most probable, are incorrect. It seems intuitive that a retrieval system could improve itself by reaching back into the lattice to find the words that were not chosen, and including them in the document. That is, the recognized speech would generate a document that included not just the most probable word, but also all other highly probable words. Surely the correct word will be in there somewhere and retrieval effectiveness will be improved.

A set of researchers at Carnegie Mellon University carried out experiments of this nature for the TREC-6 spoken document retrieval track [5, 6]. TREC is an open and competitive evaluation of a range of IR tasks, and retrieval of spoken documents was investigated from 1996 through 2000 (TREC-6 through TREC-9). The task in that year's evaluation was to retrieve the so-called "known items" within a small collection of spoken documents. Siegler et al. found that including multiple sets of hypotheses did cause some improvement, but were unable to determine the best number of hypotheses to include. They felt that it was also important to include confidence information for the additional words that were added, but did not carry out further experiments in that direction. Figure 2 shows an example of what their techniques might achieve, and also illustrates why this approach must be used cautiously: blindly, it would add several words that not only did not occur in the original document at all, but that are completely unrelated to its meaning (e.g., calm, common, and intercom).

This type of document expansion seemed promising, but has not been explored further. This failure is partly because most IR researchers are not also speech recognition experts, so do not have access to the parts of the lattice that were not provided. This encouraged the use of techniques that provide a similar effect, but that do not depend upon the inner workings of the ASR system.

3.2. Corpus-based expansion

An alternate approach to expanding documents to include words that were possibly spoken but not recognized is to use some set of expansion techniques that are common within IR research.

Query expansion is commonly used within IR to add synonyms or other related words to a query, making it more likely to find relevant documents. When the retrieval is against spoken documents, those extra words might be sufficient to compensate for the words that did not appear in the ASR output. For example, consider a hypothetical document about baseball in which for some reason the word *pitcher* was consistently misrecognized and so does not appear. If a query were issued that had the word *pitcher*, the goal of query expansion would be to add words such as *baseball*, *umpire*, and *batter*, and increase the chances that the document would be retrieved—on the assumption that it is unlikely that *all* of those words would be misrecognized consistently.

The corpus-based expansion techniques are generally referred to by names such as "local feedback" or "pseudo-relevance feedback." Although the techniques were introduced long ago [7, 8], they were not widely adopted until large corpora made the improvement they caused more likely [9]. The basic idea of all these techniques is as follows.

(1) Start with a passage of text, whether that be a query or an entire document. This is the text that will be expanded to include related words.

(2) Use that passage as a query into a collection of documents to retrieve some number of top-ranked documents. Note that a complete document can serve as a query, even though it is an unusually lengthy query.

(3) Analyze the returned set of documents to find words or phrases that occur frequently. Some approaches pay careful attention to where words or phrases occur with respect to the query words [10, 11].

(4) Add those extracted words or phrases to the query. Generally the added words are weighted less than the original query words.

As an example, consider the query "Is the disease of Poliomyelitis (polio) under control in the world?" (TREC query 302.) When expanded with a form of local feedback called local context analysis [11], some of the added words and phrases are

polio virus	dr. jona salk
polio vaccine	vaccinate programme
vaccine	polio survivor
smallpox	childhood disease
paralysis	eradicate
dr. sabin	sabin vaccine

None of those words were in the query, but they are all strongly related to the query. When the "query" is misrecognized text, it is similarly expanded to include strongly related words: many of which may help with later queries of the recognized speech.

3.3. Other approaches

Another technique that is occasionally used to retrieve ASR documents is to store phones rather than hypothesized words. Intuitively, this is something like indexing each of the sounds in a word (in order), so that "date" might be stored as *d*, *ey*, and *t*. When a query is issued, it is converted to a set of likely phones and they are searched for—for example, a query indicating that documents containing *d*, *ey*, and *t* (in that order) are likely matches.

Experiments on very small collections (300–400 short documents) suggest that this approach might be quite effective [12, 13], though it has not been widely adopted. This technique has the potential to address out of vocabulary errors in speech recognition systems, making it possible to find documents containing previously unknown or poorly recognized words. However, they will have difficulty with oddly pronounced names or other words that cannot be decomposed from text into phones easily. Further, and perhaps most importantly, it remains to be seen how well such techniques will work when scaled to substantially larger collections such as those discussed in the next section.

4. SPOKEN DOCUMENT RETRIEVAL

The problem of retrieving documents that were created by speech recognition is referred to as spoken document retrieval (SDR). In general, the queries are assumed to be typed and so without ASR errors. The queries are short in comparison to the documents, the latter of which are the output of ASR systems.

4.1. TREC SDR evaluations

The TREC SDR evaluations ran from TREC-6 in 1997 through TREC-9 in 2000. The purpose of the track was to understand the impact of ASR errors on information retrieval effectiveness. The initial expectation was that the errors would cause major problems; the reality is that SDR was not difficult to solve. Four evaluations were run.

(1) The TREC-6 SDR track [14] was a trial run to get a sense of how difficult the task was. The evaluation corpus was about 1500 stories, an incredibly small corpus in comparison to the collections of millions of documents being used for most IR research. The size of the corpus was limited by the capability of ASR systems of the time. It represents about 50 hours of speech, and ASR systems of the time ran at approximately 30 times real time, so it took about 1,500 hours to recognize just that [15]. (Modern ASR systems can run faster than real time with equivalent accuracy [16].) The task that users were given was "known item retrieval," finding a single document in response to a query. This task is substantially easier to evaluate because it does not require extensive relevance judgments: just a set of query-document pairs. The result of TREC-6 was a finding that ASR errors caused approximately a 10% drop in effectiveness, regardless of whether the queries are easy or are engineered to be "difficult" for an ASR system.

(2) In 1998 (TREC-7), the corpus was expanded to include 87 hours, or about 2900 stories. The task was changed to the more traditional document ranking task of IR: given a query, rank the spoken documents in the order they are most likely to be relevant to the query [3]. This year several speech recognition groups ran their research systems on the audio and sites had the opportunity to run the queries against a range of ASR error rates. The results made it clear that as the recognition accuracy dropped, so did retrieval effectiveness. However, the effectiveness drop was small, even when the recognition error rate reached levels of 30–40%. However, only 23 queries were used for the evaluation, making the results somewhat suspect.

(3) The third year of TREC SDR [17], TREC-8, substantially increased the size of the corpus, from less than 3000 stories to almost 22,000. This collection, derived from 500 hours of audio, was possible because ASR systems had undergone dramatic speed improvements over the few years since TREC SDR started [16]. The number of queries was raised to 50, a number that permits greater statistical confidence in the results of the evaluations [18]. Again, there were several recognition systems and sites ran cross-recognizer comparisons. The conclusion was unchanged: even against fairly high ASR error rates, retrieval effectiveness was only slightly impacted. The report from the organizers of the track [17] concluded that SDR was a success.

(4) TREC-9 saw the final experiments in the SDR track of TREC. The same corpus was used as in the previous year and the conclusions were unchanged.

SDR was a success, a non-problem. The basis of the techniques that compensated for the ASR errors was expansion.

4.2. Query expansion

In TREC-6, one site did corpus-based query expansion [19], using both a corpus of comparable documents (from the same time period as the spoken documents) as well as using the corpus of ASR documents. The following lists the expansion for one of the known item queries:

- *Original query*: What is the difference between the old style classic cinemas and the new styles of cinema we have today?
- *Basic query processing*: difference "old style" old style classic cinemas new styles cinema
- *comparable corpus expansion features*: Frankenstein "film industry" "kenneth branagh" cinema film fad style lowrie "paris cinema" "fred fuchs" "francis ford coppola" "cinemas benefit" "century rendition" "century horror classic" "adrian wootton" "art form" prod. "mary shelley" casting technician "thai house" "peter humi" profit "robert deniro" popularity "margaret lowrie" helena hollywood image
- *ASR corpus expansion features*: years trent houses emission style graduate pandering nights negotiations

cinema barrels awards kidney lott enemies "years industry" sander "houses emission" "g. o. p. fire brand set" wilderness tumor melting "majority leader trent lott" literature "cover story" dennis "house republicans" toronto soprano sequence.

The features suggested by expansion using the comparable corpus seem to be strongly related to the query: at least, they all have something to do with cinema and theater. The features selected by the same techniques from the ASR corpus (i.e., the one that will eventually be queried) look for the most part like errors.

Surprisingly, however, the known items were retrieved at the top of the list 80% of the time when the ASR features were used but only 70% of the time when the comparable features were used. It appears that expanding from the corpus to be searched added in strongly related words, and may even have added in *misrecognized* words that are consistently misrecognized across stories. Those words may serve as stand-ins for the real words that were intended.

When the same query expansion technique was applied to ranked retrieval in TREC-7 [20], the results were comparable. Retrieval effectiveness dropped about 10% with an ASR error rate of just under 30%. Effectiveness dropped uniformly as recognition errors increased, ranging from a 6% drop with a 25% error rate, a drop of 7.5% for a 33% rate, down to a 28% drop at a 66% error rate. (To ensure a fair comparison, all of those runs use the same query expansion on both the ASR text and on the human-transcribed corpus that is being used as a baseline.)

4.3. Document expansion

A potential problem with query expansion is that very short passages of text do not provide sufficient context for finding strongly related words. The longer the text is, the more likely it is that ambiguity will be reduced and semantically similar words will be included. Consider the hypothetical example above, where the word is *pitcher*. With just that single word, an expansion process might add terms related to baseball, but it might also add words related to pottery—or if the system used stemming, it might include words and phrases related to tar, the color black, or the angle of incline. However, if the starting passage of text were one or more paragraphs about baseball pitchers, the other words in the paragraph (e.g., *baseball*, *mound*, *strike*) would totally disambiguate the word *pitch* and only words related to baseball would be included.

This problem is, of course, worse when the queries being expanded are speech recognizer output. As discussed above, retrieval effectiveness dropped only slightly in SDR, even with high ASR error rates. However, when substantially shorter queries are used, the effectiveness drops almost linearly with the ASR error rates [21]. For example, 5–8 word queries with ASR error rates of 25%, 33%, and 50%, had a drop in effectiveness of 32%, 39%, and 57%, respectively. We do not know what would happen if those queries were expanded, but the lower-quality retrieval accuracy suggests that the top-ranked documents from which the expansion words would be mined would be less reliable.

For this reason, it may make the most sense to do the expansion with the full spoken document text rather than the written query. The trade offs are that the queries are accurate transcriptions but are short and may not provide sufficient context for expansion, whereas the documents are errorful, but provide substantial context for finding semantically related words and phrases.

This approach was also used in TREC-7 [22] and TREC-8 [23] and achieved excellent results. In both years, when it was compared to query expansion from several sites [20, 22], the two techniques performed essentially identically. In TREC-7, document expansion achieved a slight edge; in TREC-8, query expansion seemed to work slightly better. The queries for the SDR track were generally lengthy (about 10 words), and that may have helped avoid the problem of ambiguity in query expansion.

There were several interesting observations that came out of the SDR document expansion work [22]. First, expansion degrades if too many words or phrases are added to the document: the semantic relationship between expansion features and the original document breaks down as the number of top-ranked documents containing the term falls. Second, the expansion process must be careful about not overweighting some of the words just because they occur frequently in top-ranked documents: that is often a coincidence and does not mean that the meaning of the query should be heavily weighted in that direction. Third, it may be useful to limit expansion terms to those that were found in the speech recognizer's lattice of possibilities: that provides terms that are semantically related to the query and that also have a strong possibility of having been uttered. This last approach has the disadvantage of avoiding related words that do not appear in the speech but that might help with retrieval effectiveness. It also requires access to the speech recognizer's internals since ASR systems generally do not provide that sort of information.

4.4. SDR summary

The TREC SDR track provided a four-year venue for exploring the impact of ASR errors on document retrieval tasks. The track results were declared a success [17] because IR tasks are not very sensitive to ASR errors. Even with quite high error rates, the IR tasks had only a modest degradation in effectiveness.

The IR technique that was most successfully adopted to cope with ASR errors was expansion, either of the query or of the document. When query expansion was done, it was most successful when the expansion was done on the corpus of spoken documents—even though anecdotal evidence suggested that "better" features came from a comparable corpus. When document expansion was done, it seemed to work best when the features came from a comparable corpus.

5. TOPIC DETECTION AND TRACKING

Another research program that has investigated the impact of speech recognition errors on IR technology is topic detection and tracking (TDT). In TDT, systems are evaluated on their

ability to organize a constantly-arriving stream of news stories into groups based on the real-world events that they discuss. This involves recognizing when stories are on the same topic as well as discovering when a new topic has appeared in the news. Note that TDT does not include any query: stories are grouped automatically regardless of whether anyone is currently interested in the results.

The TDT stories come from either newswire or automatically generated transcripts of audio news for television and radio. One of the issues that TDT explores is the impact of the ASR errors on each of the tasks.[1] A task is run on speech recognizer output and compared to the effectiveness on human-generated transcripts of the audio. The latter consist either of the closed captions that come with the television audio or of closed caption-quality transcripts that were generated specifically for the TDT evaluations.

Broadly speaking, TDT researchers have found that ASR has little impact on the tasks that have a close IR parallel, and a stronger impact on the tasks that do not. Document expansion of some type is a technique that appears to help improve TDT effectiveness, both with and without ASR errors.

5.1. Some TDT tasks

We will explore three of the TDT tasks to illustrate the impact of ASR errors and how document expansion (there are no "queries" in TDT) techniques change the effect. We will talk about tracking, new event detection, and story link detection.

5.2. Tracking

TDT's tracking task starts with a small number (1–8) of on-topic stories in which a user is interested. The system's task is to monitor the stream of arriving news stories and identify which of the stories is on the same topic as the starting set. This task is strongly related to the information retrieval filtering task [24, 25] that has been explored in TREC, though filtering evaluation in that setting does not include spoken documents as a component of the problem.

Although spoken documents were an integral part of the TDT evaluation since its inception in 1998, that is also the only year that careful evaluation was done of the impact of recognition errors [26]. The reason is that it was accepted by all participants that tracking was almost entirely unaffected by ASR errors. The TDT cost became worse by anywhere from 5–10% [27] and in one case even improved [26]. In general, there appeared to be no significant difference between tracking with ASR output and tracking with closed captions. However, it was noticed that when system thresholds were lowered to create high false alarm and low miss rates, the closed caption transcripts had a substantial advantage.

That impact is shown in the curves shown in Figure 3. Those curves are taken from a run of BBN's 1998 TDT system as used in a summer workshop on TDT technology [28].

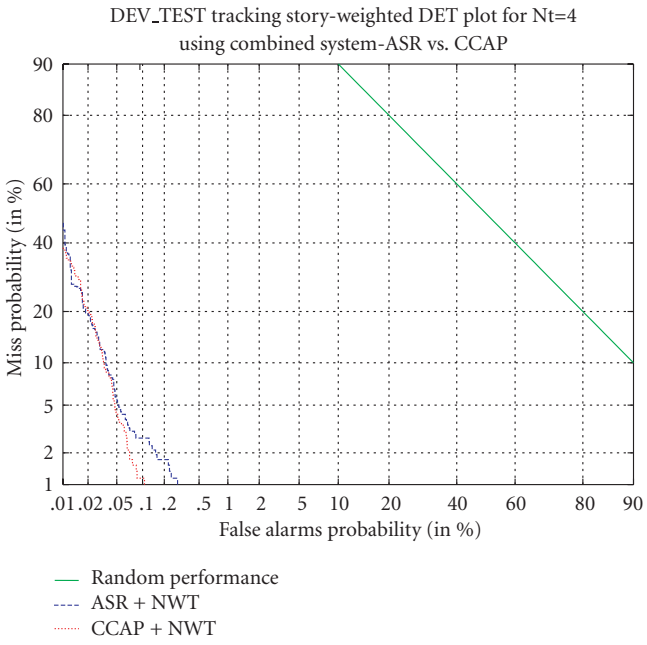

Figure 3: Plots of the 1998 BBN tracking system against a portion of the TDT development data [28]. The DET curve that is slightly to the right represents the effectiveness with spoken documents. The other DET curve corresponds to performance for closed captioning equivalents of the stories.

(The run depicted was done on development data and does not necessarily reflect BBN's official evaluation runs.) In a detection error trade off (DET) curve [29] such as that one, false alarm and miss rates are plotted against each other, with better systems moving the line toward the origin. In Figure 3, the trade off curves track almost perfectly until the miss rate is low enough, and which point the ASR system's performance degrades much more rapidly.

In the light of the SDR work discussed in Section 4, what is most surprising is that the minimal degradation with ASR documents happens even though no effort was made to compensate for ASR errors. Parameters were tuned slightly differently, but there was no substantive changes in approach. For example, the technique used by the system evaluated in Figure 3 did not include document expansion or any other attempt to ameliorate recognition errors. We hypothesize that the tracking is less sensitive to ASR errors because TDT does not have short queries: long stories provide greater redundancy and context, making it less likely that individual ASR errors will cause a problem.

5.3. New event detection

Another task within TDT is called "detection" (or "cluster detection"). The goal of the task is to group arriving stories into clusters based on the events that they discuss. All stories related to a particular event in the news should be put together—for example, stories about a particular earthquake, a specific election, or an individual crime. A key aspect to this problem is recognizing when a new event occurs,

[1] TDT stories are also multilingual, so translation issues are also central to the tasks. We ignore that issue here since every story comes with an automatically translated English equivalent.

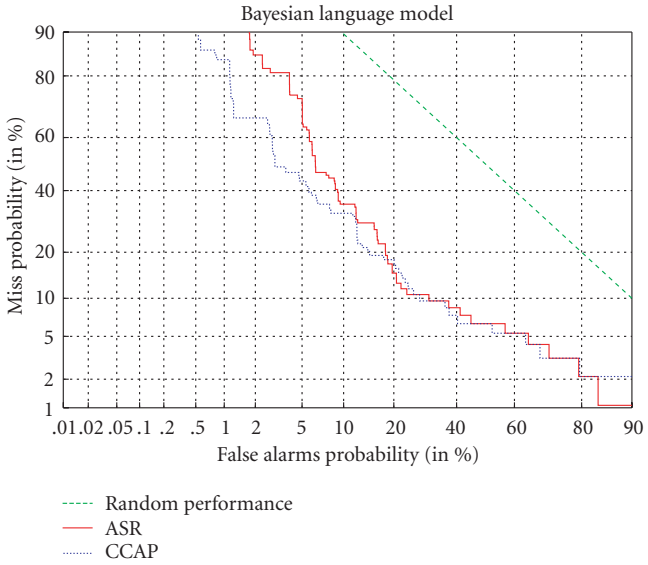

FIGURE 4: DET plot depicting contrasting performance between ASR and closed caption runs of the new event detection task. The system uses a language modeling technique to represent topics [28, 33, 34]. The dotted line (closed captions) is almost uniformly below the solid line (ASR), except in the 20–30% false alarm range.

necessitating the creation of a new cluster. The new event detection (NED) task[2] evaluates a system's ability to recognize the onset of new events.

The typical approach to NED is to build a model of all past events and then compare each arriving story to that model. If a story is sufficiently similar to one of the existing events, it is declared "not new" and is added to the mix of past events. If, on the other hand, the story is different enough from everything already seen, it is declared "new" and starts a new event in the history. Researchers have developed a range of techniques to address this problem, with the most typical being a form of vector-based document clusterings [28, 30, 31, 32].

Similarly to tracking, NED is evaluated using a cost function and a detection error trade off curve. When a system is presented with a story, if it judges it to be new but it covers an existing topic it generates a miss, and if it discusses a previously unseen topic but the system marks it as old, the system generates a false alarm. Figure 4 shows the effectiveness of a new event detection system on both closed caption and recognized documents.

The first thing to note is that the curves are substantially further from the origin: NED is a much more difficult task than tracking. The second thing to note is that the system's effectiveness on ASR documents is very much worse than that on closed caption data. For example, at a false alarm rate of 5%, the miss rate goes from about 45% for closed caption data to almost 75% for ASR documents. Neither of those is particular good results, but the impact of ASR errors is pronounced here, whereas it was minimal for tracking.

The result is initially surprising, but there is a possible explanation. Note that the curves in Figure 4 overlap strongly for false alarm rates above 20%. Recall, also, that the tracking figures in Figure 3 were almost identical for false alarm rates below 0.05% and then diverged. The sensitivity to ASR errors is flipped (with respect to error rates) in the two tasks.

This may be explained because of a strong relationship between tracking and new event detection [35]. Consider the way that most NED systems operate: they cluster all past stories and then look to see whether a new story matches any existing cluster. In some sense, this is equivalent to simultaneously *tracking* each of those clusters. If a story is not "tracked" by any of the clusters, then it is new and becomes the seed of a new cluster. If the story was, in fact, related to one of the existing clusters, then this represents a tracking miss (it should have been tracked by one of them) and a NED false alarm (it should not have been declared new). Correspondingly, if the story was "tracked" by one of the clusters but was actually on a new topic, that is a tracking false alarm (it should not have been tracked by any cluster) and a NED miss (it should have been listed as new).

This inverse relationship between errors explains why NED appears so sensitive to ASR errors while tracking did not. In fact, they are equally sensitive, but the degradation occurs in tracking at a high false alarm rate, the portion of the detection error trade off curve that is of less interest for most applications. On the other hand, for NED, the errors occur at the low false alarm region of the curve, the portion that is of greater interest. Improvements in tracking that reduce the impact of ASR errors to the right of the curve, should improve the effectiveness of NED systems. Unfortunately, there is little incentive to improve tracking at that level of errors, so little work has been done in that direction to date.

5.4. Link detection

The final TDT task that we will explore is story link detection (SLD). Unlike all other TDT tasks, SLD does not have a natural user application as an obvious extension. It is easy to imagine how someone might use a tracking system to monitor an event of interest, or how new event detection might be used as an alarm system for new events of interests. SLD, on the other hand, is a core technology rather than an application.

The job of an SLD system is to accept two random stories and to emit a decision about whether or not they discuss the same event. Note that this could clearly be used to implement a tracking system: compare incoming stories to the sample on-topic tracking stories. It could also be used to implement new event detection: when a new story arrives, if it does not discuss the same event as any prior stories, it is new. (However, it is hard to envision a user application whose sole purpose is to compare pairs of stories.)

Figure 5 shows the impact of recognition errors on the link detection task. Because SLD is a component technology for the other tasks, it is surprising how much impact ASR

[2]The new event detection task was referred to as "first story detection" for the first four years of TDT evaluation.

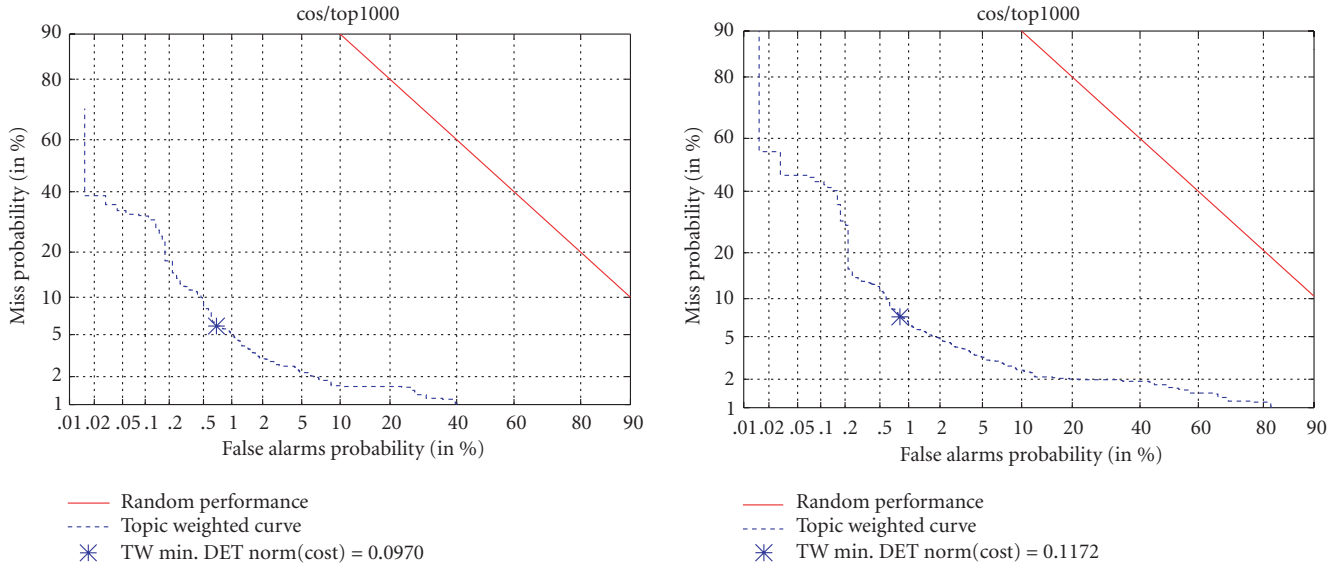

Figure 5: The impact of ASR errors on the TDT link detection task when the vector space model is used with cosine as the basis of the distance measure. The left DET graph shows effectiveness using human transcripts and the right graph shows the same for ASR transcripts. The ASR graph on the right shows an error trade off that is noticeably worse than that on the right (better curves are closer to the origin).

errors have on the effectiveness: at a 0.2% false alarm rate, the miss rate rises from about 18% to almost 30%. TDT also includes a cost function that is a linear combination of the miss and false alarm rates [36]. Following the tradition of the TDT evaluations, we show the minimum cost point on the detection error trade off curves. Here, the cost goes up (gets worse) by over 20%.

We suspect that the reason the differences are so large when they were much smaller for tracking is that SLD samples points from the entire distribution. That is, it includes story pairs that would create errors corresponding to all false alarm values on the tracking curve. So the impact of ASR is more evenly seen throughout the entire range of error values, rather than being focused on the high (for tracking) or low (for NED) false alarm range.

Figure 6 shows how the curves change when the stories are expanded using relevance models [33, 34]. That refers to a more formally justified technique for building probability distributions of words that represent the "language model" of a topic. The central idea of relevance modeling is the same as query and document expansion (build a query, find matching documents, and add features to the probability distribution). The difference is in the underlying formalisms and the motivation for carrying out the process in the first place.

In this case, the stories are each expanded using related documents from anytime earlier in the stream of arriving stories. The expanded stories are then compared and if they are similar enough, judged to be discussing the same topic. Although the change in minimum cost value is actually more dramatic than above (an increase of 35%), the detection error trade off curves are closer to each other through the entire range of errors.

Because the stories are fairly long, the redundancy of language and the occurrence of strongly related words, means that the impact of ASR errors is reduced. That is, words that are dropped by the ASR system are sometimes re-introduced to the story, and words that are mistakenly inserted by the system are de-emphasized by expanding the story with related words and phrases.

5.5. TDT summary

The Topic Detection and Tracking research program has explored the impact of speech recognition errors on its technology since its inception. The conclusion almost immediately was that ASR errors have no meaningful impact on the tracking task, but a stronger impact on other tasks. We have shown that the new event detection task is greatly affected by recognizer errors, and that there is an interesting relationship between those errors and the "less important" errors that are visible in tracking.

We have also shown in the link detection task that document expansion in TDT reduces the impact of ASR errors. Although we know of no experiments in the other tasks that explore the same issue,[3] we expect that document expansion would similarly compensate for recognition errors in those cases.

TDT includes another task called "segmentation" that requires that a system break a half hour or more of broad-

[3] Document expansion was used in the TDT segmentation task [37]. That task, not discussed in detail here, requires dividing a continuous half hour of news into topically distinct news stories. Expansion helped there, though techniques based on learning distinctive features [27, 31, 38] turned out to be more effective overall.

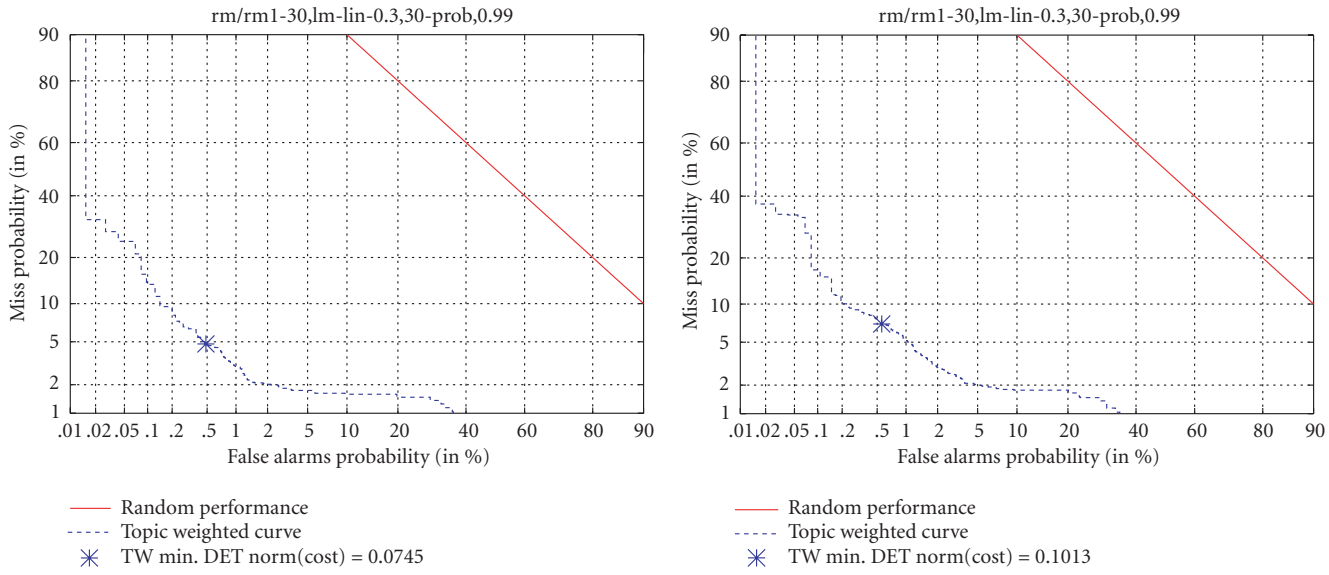

FIGURE 6: The impact of ASR errors on the TDT link detection task when relevance models are used. The left DET graph shows effectiveness using human transcripts and the right graph shows the same for ASR transcripts. Again, the ASR error trade off is noticeably worse than that using the human transcript (better curves are closer to the origin).

cast news into distinct stories. This task seems like it might be very sensitive to recognition errors, but it turns out that the cues used to signal transition between stories are robust enough to ASR errors that performance does not degrade substantially in comparison to using human generated transcripts [39]. Only a modest amount of work has been done exploring that task, so it was not included in the more detailed discussion above.

6. ROBUSTNESS BREAKS DOWN

We have shown above that speech recognition errors have only a small impact on effectiveness in SDR and TDT. For some tasks, particularly the TREC SDR task, the reduced effectiveness is almost unnoticeable, meaning that ASR can be declared essentially a non-problem [17]. It appears that ASR errors are an issue only for particular points of an accuracy trade off: the IR measures used in the TREC SDR track focus on high-accuracy portions of the trade off curve, when the false alarm rate is low. The recognition errors are more of an issue as the miss rate drops and false alarms rise.

However, IR is not just about document retrieval. There are other problems in and around IR where ASR is still likely to be a problem. To see where those are likely to be, consider any technology that works on fairly short spans of text. As mentioned in Section 4.3, ASR errors had a much more pronounced impact on effectiveness when it was short queries that were being recognized. Such technologies, when faced with ASR errors, are unlikely to find enough context and enough redundancy to compensate for the recognition failure. For such technologies, a *single* word incorrectly processed could theoretically have a profound impact.

What does this mean in terms of open problems related to speech within information retrieval systems? Here are several issues that crop up because of the length of the material being used.

• *Spoken questions of short duration*. As shown in Section 4.3, the drop in effectiveness is large for short spoken queries. How can the ASR be improved for very small snippets of speech? Is it possible for the IR system to guess that the recognition may be bad—because, for example, the query words do not make sense together? Is it just a user interface issue, where people need to be encouraged to talk longer?

• *Message-length documents*. Since short spoken items are the problem, what happens when the documents are substantially shorter? For example, voice mail messages, announcements, and so on.

• *Information extraction*. The task of identifying named entities within spoken documents is more sensitive to recognition errors. The items to be found are only a few words in length, and are often recognized by special words or grammatical structures. ASR errors introduce enough uncertainty into sentence parsing and sufficient errors into cue phrases, that information extraction is definitely hurt by ASR errors: at a recognition error rate of 40%, some types of named entity extraction operate at 20% accuracy [40] rather than their typical 90% or higher rates.

• *Question answering*. Current technologies to solve the problem of question answering (returning a specific answer to a question rather than just a document) tend to focus on small passages of text that are likely to contain the answer [41, 42]. Finding small passages in the presence of ASR errors may be an issue—and the natural language processing needed to analyze the passages may also fail.

• *Machine translation*. As more and more languages are electronically accessible, it has become important (or at least useful) to be able to look at information in multiple languages. Machine translation is currently of modest effectiveness: it is possible to get a sense of what a document from another language is about, to the extent that cross-language document retrieval is highly effective and can almost be declared a solved problem [43]. However, higher-quality machine translation—that is, that intended for a human to read—is still of dubious quality. It is unlikely that current translation technology will robustly handle recognition errors.

• *User interfaces*. Spoken documents often come grouped together (e.g., a news show with several stories) and need to be broken into segments. How can a user interface properly handle those segments, particularly when the segmentation is likely to contain errors? How can a user "skim" an audio recording to find out whether it is, indeed, relevant? It is possible to skim text, but audio must be processed linearly. Can a system provide hints to a user to help in this process?

And, of course, all of those tasks as well as the more traditional tasks of document retrieval and TDT, will suffer when recognition rates are particularly high. Speech recognition systems are reasonably accurate, but are still very error-prone in the presence of substantial background noise or conversational (nonscripted) speech [44].

7. CONCLUSION

Speech recognition errors are not a substantive problem for the traditional task of document retrieval. Given reasonably sized and accurately transcribed queries, an IR system can accurately retrieve relevant documents from spoken documents almost as well as it can from written text [17].

We believe that the reason IR tasks are not sensitive to ASR errors is that the documents being retrieved include substantial redundancy and that semantically related words also reduce the problems of having particular words lost. Both document and query expansion techniques further reduce the impact of those errors by increasing the redundancy, incorporating additional related words, as well as deemphasizing incorrectly recognized (and semantically unrelated) words.

The TDT tasks observe a greater impact from recognition errors. The tracking task appears to have little impact from ASR errors, but closer examination reveals that it is because tracking—like document retrieval—focuses on the high-accuracy end of the detection error trade off. When tasks—such as new event and story link detection—depend more on the overall accuracy, including low miss rates, ASR errors have a more pronounced impact. Although recognition errors appeared initially to be a non-problem in TDT, there is clearly room to improve over the entire set of TDT tasks.

We believe that there are substantial opportunities to investigate the impact of speech recognition errors. Some tasks, such as TDT, are sensitive to those errors in ways that were not initially obvious. Other tasks operate on very small passages of text and will necessarily be more likely to fail when individual words or phrases are corrupted.

By and large, current technologies are somewhat robust in the face of recognition errors. The degradation of their effectiveness is in proportion to the number of recognition errors: no IR-related tasks appear to fail catastrophically because of a few errors. However, the drop in effectiveness is sub-linear only for IR tasks that require very crude representations of a document. Achieving higher performance for the more fine-grained tasks is an important goal for the future.

ACKNOWLEDGMENTS

We thank Victor Lavrenko for generating the runs that were the basis of the link detection discussion. This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912. Any opinions, findings, and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

REFERENCES

[1] P. Kantor and E. M. Voorhees, "Report on the TREC-5 confusion track," in *Online Proceedings of TREC-5 (1996)*, vol. 500-238 of *NIST Special Publication*, pp. 65–74, Gaithersburg, Md, USA, 1997.

[2] V. Wu, R. Manmatha, and E. Riseman, "TextFinder: An automatic system to detect and recognize text in images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1224–1229, 1999.

[3] J. S. Garofolo, E. M. Voorhees, C. Auzanne, V. Stanford, and B. Lund, "1998 TREC-7 spoken document retrieval track overview and results," in *Proc. 7th Text REtrieval Conference (1998)*, vol. 500-242 of *NIST Special Publication*, pp. 79–89, NIST, Gaithersburg, Md, USA, 1998.

[4] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Mass, USA, 1999.

[5] M. Siegler, M. Witbrock, S. Slattery, K. Seymore, R. Jones, and A. Hauptmann, "Experiments in spoken document retrieval at CMU," in *Proc. 6th Text REtrieval Conference (1997)*, E. M. Voorhees and D. K. Harman, Eds., vol. 500-240 of *NIST Special Publication*, pp. 291–302, NIST, Gaithersburg, Md, USA, 1999.

[6] A. Hauptmann, R. Jones, K. Seymore, S. Slattery, M. Witbrock, and M. Siegler, "Experiments in information retrieval from spoken documents," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 175–181, Morgan Kaufmann Publishers, Lansdowne, Va, USA, 1998.

[7] R. Attar and A. Fraenkel, "Local feedback in full-text retrieval systems," *Journal of the ACM*, vol. 24, no. 3, pp. 397–417, 1977.

[8] W. B. Croft and D. J. Harper, "Using probabilistic models of document retrieval without relevance information," *Journal of Documentation*, vol. 35, no. 4, pp. 285–295, 1979.

[9] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic query expansion using SMART: TREC-3," in *Proc. 3rd Text REtrieval Conference*, pp. 69–80, NIST, Gaithersburg, Md, USA, 1995.

[10] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proc. 19th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 4–11, ACM Press, Zurich, Switzerland, August 1996.

[11] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Transactions on Information Systems*, vol. 18, no. 1, pp. 79–112, 2000.

[12] M. Brown, J. Foote, G. Jones, K. Spärck Jones, and S. Young, "Open-vocabulary speech indexing for voice and video mail retrieval," in *Proc. ACM Multimedia 96*, pp. 307–316, ACM Press, Boston, Mass, USA, 1996.

[13] K. Ng and V. Zue, "Phonetic recognition for spoken document retrieval," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 325–328, Seattle, Wash, USA, May 1998.

[14] J. S. Garofolo, E. M. Voorhees, V. Stanford, and K. Spärck Jones, "TREC-6 1997 spoken document retrieval track overview and results," in *Proc. 6th Text REtrieval Conference (1997)*, E. M. Voorhees and D. K. Harman, Eds., vol. 500-240 of *NIST Special Publication*, pp. 83–91, NIST, Gaithersburg, Md, USA, 1998.

[15] S. Dharanipragada, M. Franz, and S. Roukos, "Audio-indexing for broadcast news," in *Proc. 7th Text REtrieval Conference (1998)*, vol. 500-242 of *NIST Special Publication*, pp. 115–119, NIST, Gaithersburg, Md, USA, 1999.

[16] K. Lenzo, "The CMU sphinx group open source speech recognition engines," 2002, http://fife.speech.cs.cmu.edu/speech/sphinx/.

[17] J. Garofolo, C. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. 8th Text REtrieval Conference (1999)*, vol. 500-246 of *NIST Special Publication*, pp. 107–130, NIST, Gaithersburg, Md, USA, 2000.

[18] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, ACM Press, Athens, Greece, July 2000.

[19] J. Allan, J. Callan, W. B. Croft, et al., "INQUERY does battle with TREC-6," in *Proc. 6th Text REtrieval Conference (1997)*, E. M. Voorhees and D. K. Harman, Eds., vol. 500-240 of *NIST Special Publication*, pp. 169–206, NIST, Gaithersburg, Md, USA, 1999.

[20] J. Allan, J. Callan, M. Sanderson, J. Xu, and S. Wegmann, "INQUERY and TREC-7," in *Proc. 7th Text REtrieval Conference (1998)*, vol. 500-242 of *NIST Special Publication*, pp. 201–216, NIST, Gaithersburg, Md, USA, 1999.

[21] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. Kuo, "Experiments in spoken queries for document retrieval," in *Proc. Eurospeech '97*, vol. 3, pp. 1323–1326, Rhodes, Greece, September 1997.

[22] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira, "AT&T at TREC-7," in *Proc. 7th Text REtrieval Conference (1998)*, vol. 500-242 of *NIST Special Publication*, pp. 239–252, NIST, Gaithersburg, Md, USA, 1999.

[23] A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle, and F. Pereira, "AT&T at TREC-8," in *Proc. 8th Text REtrieval Conference (1999)*, vol. 500-246 of *NIST Special Publication*, pp. 317–330, NIST, Gaithersburg, Md, USA, 2000.

[24] S. Robertson and I. Soboroff, "The TREC 2001 filtering track report," in *Proc. 10th Text REtrieval Conference (TREC-2001)*, vol. 500-250 of *NIST Special Publication*, pp. 26–37, NIST, Gaithersburg, Md, USA, 2002.

[25] S. Robertson and D. A. Hull, "The TREC-9 filtering track final report," in *Proc. 9th Text REtrieval Conference*, pp. 25–40, NIST, Gaithersburg, Md, USA, 2001.

[26] J. G. Fiscus, G. Doddington, J. S. Garofolo, and A. Martin, "NIST's 1998 topic detection and tracking evaluation (TDT2)," in *Proc. DARPA Broadcast News Workshop*, pp. 19–24, Morgan Kaufmann Publishers, Herndon, Va, USA, 1999.

[27] J. Carbonell, Y. Yang, J. Lafferty, R. D. Brown, T. Pierce, and X. Liu, "CMU report on TDT-2: Segmentation, detection and tracking," in *Proc. DARPA Broadcast News Workshop*, pp. 117–120, Morgan Kaufmann Publishers, Herndon, Va, USA, 1999.

[28] J. Allan, H. Jin, M. Rajman, et al., "Topic-based novelty detection: 1999 summer workshop at CLSP, final report," Tech. Rep., The Johns Hopkins University, Baltimore, Md, USA, 1999.

[29] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech '97*, vol. 4, pp. 1895–1898, Rhodes, Greece, September 1997.

[30] R. Papka and J. Allan, *Topic Detection and Tracking: Event Clustering as a Basis for First Story Detection*, pp. 96–126, Kluwer Academic, Boston, Mass, USA, 2000, Advances in Information Retrieval: Recent Research from the CIIR.

[31] Y. Yang, J. Carbonell, R. Brown, J. Lafferty, T. Pierce, and T. Ault, "Multi-strategy learning for TDT," in *Topic Detection and Tracking: Event-Based Information Organization*, J. Allan, Ed., pp. 85–114, Kluwer Academic, Boston, Mass, USA, 2002.

[32] J. Allan, V. Lavrenko, and R. Swan, "Explorations within topic tracking and detection," in *Topic Detection and Tracking: Event-Based Information Organization*, J. Allan, Ed., pp. 197–224, Kluwer Academic, Boston, Mass, USA, 2002.

[33] V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard, and S. Thomas, "Relevance models for topic detection and tracking," in *Proc. Human Language Technology Conference (HLT 2002)*, pp. 104–110, San Diego, Calif, USA, March 2002.

[34] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127, New Orleans, La, USA, September 2001.

[35] J. Allan, V. Lavrenko, and H. Jin, "First story detection in TDT is hard," in *Proc. 9th International Conference on Information and Knowledge Management*, pp. 374–381, McLean, Va, USA, November 2000.

[36] J. G. Fiscus and G. Doddington, "Topic detection and tracking evaluation overview," in *Topic Detection and Tracking: Event-Based Information Organization*, J. Allan, Ed., pp. 17–31, Kluwer Academic, Boston, Mass, USA, 2002.

[37] J. Ponte and W. B. Croft, "Text segmentation by topic," in *Proc. 1st European Conference on Research and Advanced Technology for Digital Libraries*, pp. 113–125, Pisa, Italy, September 1997.

[38] D. Beeferman, A. Berger, and J. Lafferty, "Text segmentation using exponential models," in *Proc. 2nd Conference on Empirical Methods in NLP*, Providence, RI, USA, August 1997.

[39] M. Franz, J. McCarley, S. Roukos, T. Ward, and W.-J. Zhu, "Story segmentation with hybrid models," 1999, Presentation available at http://www.nist.gov/TDT/tdt99/presentations/.

[40] V. Goel and W. Byrne, "Task dependent loss functions in speech recognition: Application to named entity extraction," in *ESCA ETRW Workshop on Accessing Information in Spoken Audio*, pp. 49–53, Cambridge, UK, 1999.

[41] J. Prager, J. Chu-Carroll, and K. Czuba, "Statistical answer-type identification in open-domain question answering," in *Proc. Human Language Technology Conference (HLT 2002)*, pp. 137–143, San Diego, Calif, USA, March 2002.

[42] E. M. Voorhees, "Overview of the TREC-9 question answering track," in *Proc. 9th Text REtrieval Conference*, pp. 71–80, NIST, Gaithersburg, Md, USA, 2001.

[43] J. Xu and R. Weischedel, "TREC-9 cross-lingual retrieval at BBN," in *Proc. 9th Text REtrieval Conference*, pp. 106–115, NIST, Gaithersburg, Md, USA, 2001.

[44] T. H. Crystal, A. Schmidt-Nielsen, and E. Marsh, "Speech in noice environments (SPINE) adds new dimension to speech recognition R&D," in *Proc. Human Language Technology Conference (HLT 2002)*, pp. 200–207, San Diego, Calif, USA, March 2002.

James Allan is an Assistant Professor in the Department of Computer Science at the University of Massachusetts Amherst, where he also codirects the Center for Intelligent Information Retrieval. His research interests include the automatic indexing, organization, and retrieval of text. One focus of his work is topic detection and tracking (TDT), where the domain is written and spoken broadcast news stories. Allan received the Ph.D. degree from Cornell University in 1995.