Passage Feedback for News Tracking

Hema Raghavan and James Allan
Center for Intelligent Information Retrieval,
140 Governor's Drive,
University of Massachusetts,
Amherst, MA -01003
{hema,allan}@cs.umass.edu

ABSTRACT

We extend the event tracking task of Topic Detection and Tracking (TDT) to create a framework in which a user can highlight relevant passages in addition to specifying the relevance of documents. A dual framework of combined document and passage feedback improves performance over a state-of-the-art system without feedback by over 70% . Although annotators vary in the content and length of the passages marked for feedback, improvements in performance are consistent. We demonstrate how events in news follow certain trends over time, making passage feedback critical to event tracking.

1. INTRODUCTION

With ease of access to online editions of almost all major newspapers (148 of the 150 of the leading newspapers in the United States were online by 2003 [19]), people have access to vast amounts of data. People are increasingly using the internet as their primary source of news [4, 16]. A 2005 survey[16] shows that there are about 39.3 million unique visitors to the top 10 newspaper websites in the United States. With print newspapers, people got all their news from a few sources, effectively allowing the editors of those sources to do the filtering of content for them. The availability of so many online sources provides a person greater choice, with multiple perspectives and levels of detail for the same news story.

The online medium also facilitates personalization. A user can specify a topic of interest to a filtering or a tracking system by means of a few keywords, like a query to a search engine. The system delivers new stories on the specified topic as and when they appear in the news. The user may provide some form of relevance feedback as stories are delivered .

Document filtering has been considered as a research problem by the TDT and TREC communities[2, 21]. In this paper we consider the problem of tracking events in news as studied by the TDT community. The system is given one example document per topic (or event) in place of a query and the test stories arrive in a stream. The system is expected to filter the on-topic stories from the off-topic ones. We use the terms tracking and filtering interchangeably in this paper.

The TDT supervised tracking task[13] and the TREC adaptive filtering task[21] consider a scenario where the user is willing to provide feedback on the relevance of documents. We find that asking the user to highlight terms, passages or phrases for documents that they judge relevant can significantly boost performance over using just document feedback.

We describe our experimental setup to measure the effectiveness of document and passage feedback in sections 3, 4 and 5. We discuss our results which show that passage feedback consistently outperforms document based feedback in section 6. These results are significant, with passage feedback on as few as 20 documents resulting in more than 70% improvement in cost over a traditional TDT unsupervised tracking system. Additionally, we plot performance as a function of the number of feedback iterations in our evaluations and find that most of the gains are in the initial feedback iterations. This result allows us to remove the unrealistic assumption of TDT and TREC that the user is willing to provide feedback on every delivered document.

In section 6.1 we explain the above results (the marked improvement in performance and the lack of need for feedback after 20 iterations) by showing that the way events are discussed in news follows certain trends over time, making passage feedback useful to event tracking. We also study how the terms used to describe an event evolve in section 6.2. We find that most of the important keywords which describe the event appear in the first few stories itself. User interaction in identifying these keywords gives significant gains in performance.

The goal of this work is neither to highlight the importance of passages in general, nor to propose methods for automatic passage retrieval. Rather, the aim is to highlight the importance of sub-document level feedback for the specific task of news filtering. The domain here is news, the task is filtering and the entities of interest are events. Events in news are usually bursty: an event occurs, there is an increased interest in the event with many articles on the topic, and eventually interest on the topic fades out. For some anticipated events like a predicted hurricane or an expected conference, interest in the event actually builds up slowly over time, with interest peaking when the event actually happens. Evolution of news events has been studied before [14, 8, 20]. However, all that work has focused on how new terms appear in news over time in the corpus. Our work specifically considers how documents on a given topic evolve over time. Our work is the first to bring to light why and how the evolution of events motivates passage based feedback.

We also discuss how people vary in the quantity and content of text that they mark relevant in section 6.3. Our two annotators overlap in the text they highlight only 65% of the time. In spite of these

Submission for SIGIR 2006. Please do not cite or distribute this draft paper.

variations, performance improvements are significant for both of them.

We also measure the performance of our systems using the TREC Utility Metric (T11SU) in section 7. There has been little work on finding systems that do well on both the TREC and TDT utility measures simultaneously[24]. Passage feedback gives significant improvements on both these metrics.

For "noisy" documents like the output of machine translation, we find that people can determine relevant passages although it is difficult to understand the entire document. We show how these highlighted passages can be used to improve the accuracy of multilingual tracking in section 8. We now move on to discuss our work in context with other related work.

2. RELATED WORK

Our work is directly related to TDT tracking and TREC filtering. It is also related to passage feedback (both interactive and automatic) which has been well studied in information retrieval.

The Information Retrieval community has been studying the problems associated with tracking or filtering stories from multiple news sources for many years through the TREC filtering [21] and TDT [13] evaluations. The two evaluations differ in the kinds of topics that are tracked. Although TREC has used news corpora, the emphasis has not been on news per se. TREC topics are more subject based, e.g., *the effects of osteoporosis*. TDT research has concentrated on news and the topics to be tracked are events, for example *Hurricane Mitch, The USS Cole Bombing* and so on.

In TREC filtering, the topic is specified by means of a few keywords and the filtering system is expected to detect documents on the topic of interest in the incoming news stream. In TDT, the topic to be tracked is determined by one sample training story.

The TDT tracking task has largely been unsupervised, where no interaction is permitted after the initial training document is provided. The TREC filtering task has always been supervised. The task assumes that feedback is provided on every delivered document. TDT conducted a supervised adaptation track in 2004 where TREC-style user feedback was permitted [13]. Improvements in performance were found with the introduction of interaction.

Supervised adaptation for TDT tracking has also been studied in the past by Leuski and Allan [12]. They highlight the unrealistic assumption of supervised filtering that the user is willing to provide feedback on every single document that is delivered. That work did not consider sub-document feedback. In this paper we assume that we have a budget on user interaction time, and because of this budget we want that the information we get from the user at each iteration of interaction should result in a significant decrease of system error. We find that sub-document feedback gives big improvements in performance at each iteration of interaction.

The TDT and TREC tasks differ in their evaluation metrics. The TDT task is more recall oriented (measured by cost) and the TREC task gives greater emphasis to precision (measured by T11SU). A recent paper by Yang et al [24] highlights the differences between the two. In this paper we consider both evaluation metrics and find that passage feedback results in improvements in performance.

Giving a human control over what passages must be fed back has been studied in interactive information retrieval with mixed results [23, 7]. These systems also allow the human to control the weights of feedback terms or to specify complex relationships (like hypernyms and synonyms) between the query and feedback terms. In this work we ask the annotator to simply highlight relevant passages ignoring the underlying system. Contextual search by Yahoo![1] where a user can highlight a portion of a document while reading it, is also similar in intuition.

Passage feedback was found useful for the TREC routing task since long documents contain a lot of "noise" that is harmful for feedback [1]. In that work the feedback passages were automatically determined from the relevant documents. The importance of evidence from passages for information retrieval is also well known and therefore automatic passage retrieval has been well studied [18, 5].

3. DATA

For experiments in this paper we used the TDT3 and TDT4 corpora [13] provided by the Linguistic Data Consortium (LDC) and used in the TDT evaluation runs from 1999 to 2003. The corpus consists of data from newswire and broadcast (radio and TV) sources in English, Arabic and Mandarin. Output of an automatic speech recognition (ASR) system is made available for the broadcast sources. Machine translated (MT) text is available for the non-English stories. The TDT3 corpus ranges from October though December 1998 and the TDT4 corpus spans October 2000 through January 2001. The corpus is annotated for topics which are events in news.

Details about the corpora and the topics are available on the LDC webpage [13]. Since ASR and MT output are difficult to read, we obtained passage level relevance judgments for 90 topics (30 from TDT3 and 60 from TDT4) on English newswire documents only. We also obtained limited judgments for machine translated text. We discuss experimental results on that corpus in section 8.

One annotator (the first author of this paper) did passage level judgments for all 90 topics. This annotator was shown the description provided by the LDC for each topic: a summary and a list of people, places and organizations that are key to the topic. For each topic, the annotator looked at all *relevant documents* (as determined by the LDC) in chronological order. The annotator was asked to skim the document and mark passages that made it clear that this document was indeed relevant to the topic. Unlike experiments in interactive retrieval where the user is aware that the document or passage that they mark will be fed back to a retrieval system, the judgments we obtained asked the user to mark sections of text that determine relevance in a manner akin to browsing a book to highlight relevant passages. The annotator was not concerned about the underlying system or task. Passages were allowed to be as short as five characters with no upper limit on the length, giving the annotator the flexibility to mark terms, phrases or entire documents. For example, in the case of topics like *Hurricane Mitch* the occurrence of the phrase "hurricane Mitch" is a sufficient indicator of relevance for most humans. On the other hand in some cases, like for very short documents, the entire document may be relevant. The annotator kept in mind that speed rather than thoroughness was important.

The annotator also indicated the "degree of relevance" of a document. There were four choices in this regard. A document was considered *fully relevant* if it almost entirely (for more than about 70% of its text) talked about the topic. A document was *partially relevant* if large parts of the document were not exactly on topic. For example, documents on the merger of Chrysler and Daimler-Benz also refer to other recent mergers in the automobile industry. The annotator had the option of marking the degree of relevance as *don't know* when confused. There was also a fourth option of *non-relevant*. Most documents were found to be fully relevant (74.72%) and about a quarter (23.8%) were found partly relevant. Only 0.1% of the stories were judged as *don't know*. No stories were marked non-relevant; the annotator never found anything that the LDC had marked relevant to be non-relevant.

[1] http://yq.search.yahoo.com

Since the documents were shown to the annotator in the order of their appearance in the news, the evolution of a topic and the changing awareness of the annotator are also implicitly modeled in the annotations. The annotator was asked not to bother repairing previous relevance judgments in cases where increased awareness revealed a past error. For example, in the case of the *USS Cole Bombing*, mentions of the indictment of Osama Bin Laden at first seem speculative, but as more and more news agencies report it, the passages talking about Bin Laden's involvement in the bombing seems relevant to the topic.

To see whether a non-expert – i.e., a person who had not necessarily worked in TDT or TREC filtering would be able to highlight passages with the same effectiveness as the author-annotator, we did additional experiments to study inter-annotator agreement in section 6.3.

We measured the time taken by the annotator to mark relevant passages, i.e., the time taken from when the document was shown to her to when she clicked on the *save* button. We explicitly told the user that they were being timed, and to click on *coffee break* if they were going to pause the annotation. We found the median time to mark a document for relevant passages to be 29 seconds (average time is 50s). We report the median to remove the few outliers where the user forgot to use the *coffee break* option, thereby inflating the average. We did not measure the time it would have taken to simply mark the relevance of a document without having to highlight passages.

4. EVALUATION

In the official TDT evaluation [2] task the system is given one training document per topic. For each topic, the test data consists of a stream of documents that arrive in chronological order and need to be declared as on or off the topic of the training story. The system is expected to output a YES/NO decision for each story in the stream along with a confidence score. The task is online, i.e., the system has to process the test stream in order and no look-ahead is allowed. In the unsupervised tracking task no feedback is allowed after the initial training document is provided. In the supervised adaptation track the user provides a relevance judgment on every delivered document. This can then be used by the system in order to adapt itself.

In both tasks, a test document is compared to a model of the topic. If the similarity exceeds a given threshold, a YES is output, otherwise a NO is output. The confidence score (typically a similarity score) is also output by the system.

Based on the number of documents on topic that the system failed to deliver and the number of off-topic documents that the system delivered, the probability of a *miss* (P_{miss}) and a *false alarm* (P_{fa}) at a given threshold can be computed as:

$$\begin{array}{lcl} P_{\text{miss}} & = & \textit{\#miss/\#Targets}, \\ P_{\textit{false alarms}} & = & \textit{\#fa/\#NonTargets}. \end{array}$$

#miss and *#fa* are the number of misses and false alarms respectively. The stream comprises of *#Targets* and *#NonTargets* on-topic and off-topic stories respectively. The *cost* is a linear combination of the misses and the false alarms and is given as:

$$\begin{array}{rcl} C_{\text{DET}} & = & C_{\text{miss}} * P_{\text{Target}} * P_{\text{miss}} \\ & + & C_{\text{fa}} * (1 - P_{\text{Target}}) * P_{\text{fa}} \end{array}$$

The cost of a miss (C_{miss}) and false-alarm (C_{fa}) are typically set to 10 and 1 respectively, and P_{Target} is apriori determined to be 0.02. The normalized cost ($cost$) is computed by dividing C_{DET} by $Min(C_{\text{miss}} * P_{\text{Target}}, C_{\text{fa}} * (1 - P_{\text{target}}))$. The cost can be greater than 1. With the current choices of C_{miss} and P_{Target}, the cost is 1 if the system outputs a NO decision for all documents in the stream and zero for no misses and no false alarms. The lower the cost, the better the system. Costs are averaged over topics so that very large topics do not skew the score. P_{miss} and P_{fa} for all decision thresholds are plotted as a DET plot (or ROC curve). The minimmum value of the normalized cost function (min-cost) gives the performance at the best threshold. The DET plot (or ROC curve) and the min-cost give a measure of the quality of the ranking.

In our evaluations we consider that there is a budget on the users time and the user is willing to provide feedback only n times. We therefore, plot cost for increasing values of n.

Although the TDT cost function is our main evaluation metric, in section 7 we report results on the TREC supervised filtering metric as well. We describe that measure at that time.

5. SYSTEM DESCRIPTIONS

For each topic we now have the complete list of relevant documents, and relevant passages within these documents. The documents that are not judged are considered off-topic. Given these relevance judgments we now proceed to simulate supervised filtering in the style of TREC and TDT, where there is one training document for each topic and the test data for each topic is a stream of documents as described in section 4.

All our systems used the vector space model with feedback following a Rocchio [17] like framework which has been found to be effective for supervised filtering at TREC [22] and TDT [13]. As mentioned earlier, Yang et al[24] discuss the similarities and differences between TREC and TDT filtering. One of the aspects of the TDT cost function that they point out is that it is highly recall oriented as compared to the TREC metric. For the TDT measure they find that a Rocchio based system[17] works very well, whereas a logistic regression based system is the best for the TREC evaluation measure. Since the primary evaluation measure is the TDT cost function, we use a vector space model.

Let the vector corresponding to the initial training document be denoted as D_0. Let P_0 denote the vector corresponding to the set of words highlighted in D_0 (i.e., the passages). D_0 is weighted by term frequency and truncated such that only the top 50 terms are retained. Similarly, the terms in P_0 are the terms in the highlighted passages of D_0 and the weights are the frequencies of these terms in the document (instead of passages, since we found this method to work better). D_0 and P_0 are used to create two topic models: a document model, $M_{D_0} = D_0$ and a passage model, $M_{P_0} = P_0$. The test documents $D_1 ... D_T$ are processed in order. A vector P_t is constructed for each document D_t in the same way that P_0 was constructed from D_0. At each iteration, t, the document D_t is compared to the topic models from the previous iteration, $M_{D_{t-1}}$ and $P_{D_{t-1}}$, using the cosine similarity metric, *cos*:

$$\begin{array}{rcl} \textit{confidence}(D_t) & = & \lambda \, cos(D_t, M_{D_{t-1}}) \\ & + & (1 - \lambda) \, cos(D_t, M_{P_{t-1}}) \end{array}$$

A YES is reported (i.e., a document is delivered) if the *confidence* score is above a threshold d. If the *confidence* is very high (as determined by some predefined threshold c) then feedback is not asked, so as to minimize the amount of interaction for the user. If the document lies in an *uncertain range* ($d < \textit{confidence} < c$) and the budget ($n$, the number of times feedback may be asked of the user) is not exhausted, feedback is asked. An iteration in which feedback is obtained is called a feedback iteration. We ask for feedback only

Name	Training	**Feedback**	λ
1. D-D	$M_{D_0} = D_0$	$M_{D_t} = M_{D_{t-1}} + D_t$	$\lambda = 1$
2. P-P	$M_{P_0} = P_0$	$M_{P_t} = M_{P_{t-1}} + P_t$	$\lambda = 0$
3. DP-D	$M_{D_0} = D_0$ $M_{P_0} = P_0$	$M_{D_t} = M_{D_{t-1}} + D_t$ $M_{P_t} = M_{P_0}$	$0 < \lambda < 1$
4. DP-P	$M_{D_0} = D_0$ $M_{P_0} = P_0$	$M_{D_t} = M_{D_0}$ $M_{D_t} = M_{P_{t-1}} + P_t$	$0 < \lambda < 1$
5. DP-DP	$M_{D_0} = D_0$ $M_{P_0} = P_0$	$M_{D_t} = M_{D_{t-1}} + D_t$ $M_{P_t} = M_{P_{t-1}} + P_t$	$0 < \lambda < 1$

Table 1: A system named X-Y trains on X and uses Y to update its model, where possibilities are D for the full document, P for the highlighted passage, and DP for a blending of the two.

on uncertain documents because we want to consider maximizing performance for budgets of n feedback iterations. In other words, we do not want to ask a question whose answer we are fairly confident about. During feedback if the document is relevant, then M_{D_t} and P_{D_t} are updated as:

$$\begin{aligned} M_{D_t} &= M_{D_{t-1}} + D_t \\ M_{P_t} &= M_{P_{t-1}} + P_t \end{aligned}$$

If no feedback is obtained, then $M_{D_t} = M_{D_{t-1}}$ and $M_{P_t} = M_{P_{t-1}}$.

We could have used pseudo relevance feedback to adapt the document model with documents having a *confidence* greater than c, but we found no empirical advantage to doing that on our training set. It also creates the possibility of adapting to false alarms.

We consider several combinations of document and passage feedback, resulting in the five systems shown in table 1. The D-D system does not include any passage feedback. This system resembles the typical supervised adaptation framework of TREC and TDT, except that feedback is provided only on uncertain documents and not on all delivered documents. The P-P system has $\lambda = 0$ implying that only the highlighted passages are used in all iterations. The DP-D system has passage level markings only for the initial training document, D_0. System DP-P is trained with the document and the passage but has only passage feedback. System DP-DP uses passages and documents for both training and feedback. Systems DP-D and DP-P help measure how much each of document and passage feedback contribute to system DP-DP.

All systems were trained for d, c and λ on a randomly picked subset of 5 topics and tested on the remaining 85 topics. The optimal values of parameters were found to be $c = 0.4$ and $d = 0.15$ (all systems) and $\lambda = 0.5$ (systems 3, 4 and 5).

6. RESULTS AND DISCUSSION

The case when $n = 0$ corresponds to the TDT unsupervised tracking task where the system receives no feedback after the initial training document is provided. The DET curves (that form part of the traditional TDT evaluation) for the D- and DP- system when $n = 0$ are shown in figures 1(a) and 1(c). Passage feedback is consistently better than document feedback.

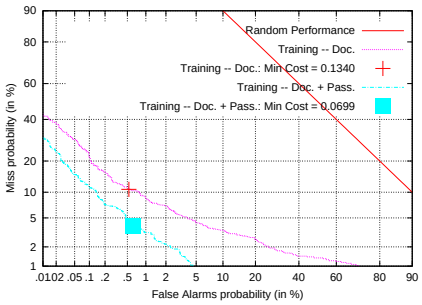
The cost and the minimum cost obtained by our systems for increasing number of iterations of feedback (n) is shown in figure 1. To get some intuition for what the cost function means note that the cost and the miss rate are related as *cost* $\leq P_{miss}$ (The cost equation works out to $C_{Norm} = P_{miss} + 4.9P_{fa}$). The equality is reached when *#fa=0*. From the very low costs in figure 1 we can see how few documents are missed.

As the number of feedback iterations increases, all systems improve in performance, with the systems that incorporate passage feedback showing the most rapid changes. Performance improvements are significant in the early feedback iterations (n), but it saturates very quickly. This has been observed in TDT before [3]. In general, systems DP-P and DP-DP show statistically significant (using a two tailed t-test with 95% confidence) improvements over the baseline system D-D on cost and min-cost at $n = 20$ for both corpora. Thus passage feedback by itself improves performance over document feedback. Mixing document with passages improves performance over using only passages. System D-D at $n = 0$ has a cost of about 0.17, and system DP-DP has a cost of about 0.05 at $n = 20$; over 70% drop in cost.

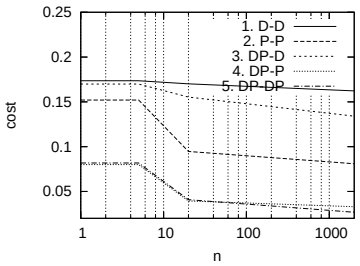
Using only documents for feedback (Systems D-D and DP-D) does not seem to improve performance on the TDT3 corpus, as n grows. On the TDT4 corpus however, the improvements are as expected (figure 1). The result for the TDT3 corpus seems to counter what we know from information retrieval [9] – that relevance feedback results in performance improvement. It also seems contrary to previous work [12] that found performance improvements in the TDT3 corpus with increasing numbers of feedback documents. That work used the entire TDT3 corpus and it is possible that the English newswire subset that we used is not really benefited by the introduction of document feedback.

6.1 Trends in Relevance

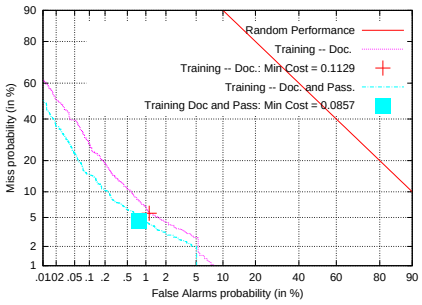
Figure 2 shows a plot of the average degree of relevance (see section 3) of the relevant documents when ordered by their time of appearance in the news. *Partially relevant* documents receive a score of 2, *fully relevant* documents receive a score of 3. Documents marked as *don't know* and *non-relevant* receive scores of 1 and 0 respectively. For example, in figure 2(a) the first relevant document has a score of 3, the second has a score of 2 and so on. Figure 2(b) plots the average relevance of the first 100 relevant stories. The average is computed over all topics. There are several documents in the beginning that are only partially relevant. For example, topic 40059 ("The bombing of the USS Cole") has some partially relevant documents in the beginning that mainly discuss the Bush-Gore campaign and foreign policy in the context of the then-recent bombing of the USS Cole. The partially relevant stories towards the end occur many months later, and largely discuss terrorism, citing references to the Cole bombing that had happened earlier that year. This trend explains the fluctuations we get with system DP-D. Adapting to partially relevant documents in the beginning introduces noise into the topic models, worsening performance. The fully relevant documents occurring after the initial set of partially relevant documents are probably never detected and therefore the model is not able to recover from its mistakes. The passage based topic model (M_{P_t}) is always topically cohesive with little noise. Using document and passage models in tandem works the best of all because there are useful cues in the words that occur in the document (words like *terrorism* and *middle east* in stories on the USS



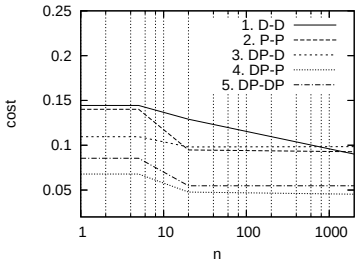
(a) TDT3 DET Plot at $n = 0$



(b) TDT3 cost

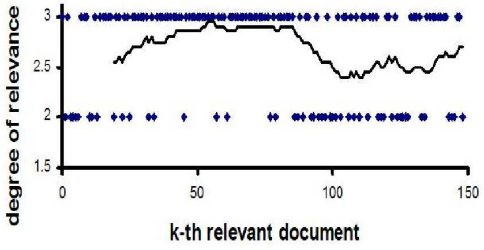


(c) TDT4 DET Plot at $n = 0$

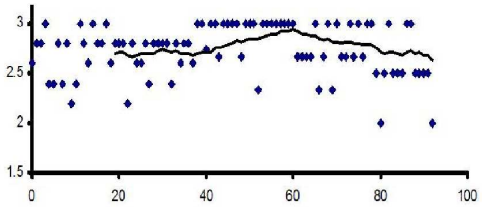


(d) TDT4 Cost

Figure 1: Cost and min-cost for increasing feedback iterations, n. The rightmost column is the det-plot at $n = 0$ (no adaptation). The det-plots show the DP- and D- models. The upper and lower row correspond to the TDT3 & TDT4 corpus respectively. The cost and min-cost are shown at values of n. The lower the cost, better the performance.

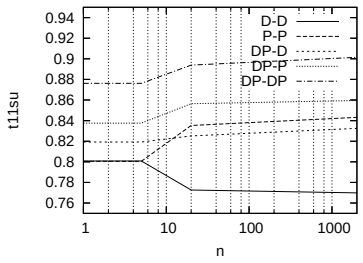


(a) Topic 40059

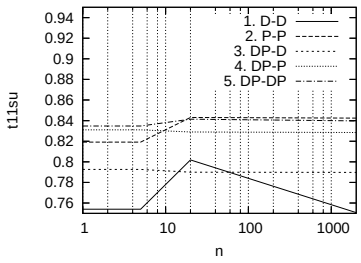


(b) Avg. degree of relevance

Figure 2: Trends in relevance: The "degree of relevance" of the relevant documents when sorted by time. The x-axis is the k-th relevant document. Figures 2(a) shows the trends in relevance for one example topic. Figure 2(b) shows the average trend over all topics. Moving averages with a window size of 20 are shown. Observe how interest in a news event builds up over time and then drops off.



(a) TDT3



(b) TDT4

Figure 3: T11SU for increasing feedback iterations, n. The higher the T11SU values, better the performance.

Cole bombing) that can be leveraged to improve performance.

6.2 Informative terms in news

From figure 4 it seems like after $n = 20$ feedback iterations there is little gain in engaging the user in further feedback. We now analyze why this is the case. Our hypothesis is that there are only a few important keywords that describe a news topic. Almost all news stories on a topic, whether they are partially relevant to a topic or fully relevant will at some point in the story relate it to the main thread of the topic using a certain small set of keywords. For example, stories on the USS Cole Bombing, will almost always contain the words *USS, Cole, Yemen, ship* etc. It is possible that our annotators are able to highlight these important keywords in the first few documents itself and most of the gain lies in being able to find these keywords. On the other hand, a system that relies only on document level feedback needs to see many more documents on the topic of interest before it identifies this small set of discriminatory words. Since we know that many documents are only partially relevant, automatically determining the useful keywords from only documents may be difficult.

In this section we investigate the evolution of important keywords in news topics. Towards this goal we design the following experiment. We measure the importance of a term to a topic by computing its information gain score. Information gain [6] is given as:

$$IG = \sum_{c \in \{-1,+1\}} \sum_{\tau \in \{0,1\}} P(c,\tau) log \frac{P(c,\tau)}{P(c)P(\tau)}$$

where τ is 0 or 1 indicating the presence or absence of a feature in an on topic ($c = +1$) or off-topic ($c = -1$) document respectively.

We can compute a list of information gain scores for all terms in the vocabulary of the corpus for each topic. The most informative keywords would have the highest score. Now, for each topic, we order the documents by their time of appearance in the news. Thus the i^{th} document appears after the $(i-1)^{th}$ document. Then proceeding in the order of time, for each document we sum the information gain scores of terms that first appear in that document. In this way, each document gets an *informativeness score*. Note that if a term has already occurred in a previous document on that topic, it does not contribute to the informativeness score. The informativeness score measures how many new keywords appear in each document for a given topic. We can thus plot the informativeness scores over time for each topic. Since information gain scores are not normalized, we need to normalize the informativeness score for each topic by dividing by the maximum for that topic. A normalized informativeness score of 1 then represents the most informative document for that topic.

Figure 4 shows the normalized informativeness score of the i^{th} document on a topic (sorted by the time at which the news story appeared) for 10 topics of the TDT4 corpus. The line represents the average informativeness score of the i^{th} document on a topic.

The informativeness score of a document measures how many important new terms are present in the document. From the plot in figure 4 it seems like the key informative terms appear in the first 20 documents. Note that this is not counter-intuitive to the discussion on the degree of relevance of a document. That is, a document that is partially relevant will still use the key descriptive words of the topic in the portions that refer to the topic. The passage containing these key-words thus exists in these partially relevant documents amidst other passages that are only tangentially related to the topic.

6.3 Inter-Annotator Agreement

We asked a second annotator to mark 10 randomly chosen topics

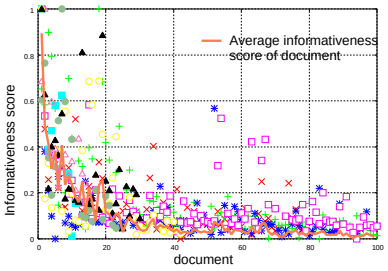


Figure 4: The evolution of terms with time: The informativeness scores of the i^{th} document on a topic. The x-axis is the i^{th} document. The y-axis is the informativeness score. The thick line traces the average informativeness score over all topics. Most of the informative terms appear in the first 20 documents.

(465 documents) to test how much annotators vary in their choice of relevant passages. This annotator was not a computer scientist and was not made to understand the underlying task. We asked this person to skim through documents, highlighting text that indicated that the document was on the topic of interest as described in section 3. We call this subset of the corpus, the TDT3′ corpus.

Evaluations for summarization [10] have found that when humans are asked to mark the most important sentences in a passage to generate 10% summaries, they are fairly consistent. But as the length of the summary is increased from 10% to 20%, humans often disagree and there is a large amount of variance in what they mark. The disagreement and variance also depend on the underlying corpus and task.

The annotators agreed about the degree of relevance of a document about 76% of the time (Kappa=0.3), which can be interpreted as a "fair" amount of agreement [15]. LDC used to mark "briefly" relevant stories (less than 10% relevant) in the TDT corpus. They found that annotators often disagreed on *relevant* versus *briefly relevant* stories [2]. We cannot use the Kappa-measure to measure inter-annotator agreement on the passages because we have no notion of non-relevance. Nor can we use percentage agreement which is commonly used for summarization evaluations as our annotations were not limited to marking complete sentences nor was a limit on the length of the passage imposed. Hence, we measured how much the two annotators agreed as follows.

Let S_i represent the set of terms in all relevant passages marked by the first annotator in document i. Let S'_i be the corresponding set for the second annotator. We measured inter-annotator agreement as $(S_i \cap S'_i)/(min(|S_i|, |S'_i|))$, which measures how much the two annotators agreed relative to the more terse one. Using this measure we found that our annotators agreed for about 65% of the terms. Because of the high variance in annotator judgments we report results for the two annotators separately.

Figure 5 shows the cost obtained by both the annotators on the TDT3′ topics using the DP-P system. We compare our users on the DP-P system since it measures the quality of the passages fed back in the feedback iterations. In spite of differences in annotation, the graphs are very similar, which shows that passages marked by both annotators were beneficial. We believe that whatever bias was caused by having a single annotator for most evaluations does not significantly affect our results.

We measured the time taken to annotate passages for the second annotator also. We found she took a median of 43 seconds (aver-

[2] http://www.ldc.upenn.edu/Projects/TDT3/email/email_360.html

age 57s) to mark documents for passage level annotations. This is slightly higher (1.5 times) than the time the author annotator took and has to do with the fact that the author-annotator probably understood better that the passage markings could be fairly crude. We believe the second annotator was more careful.

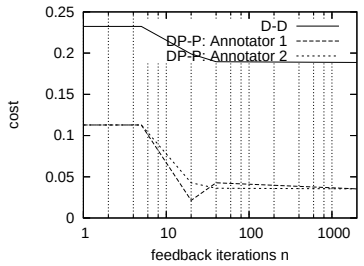


Figure 5: Inter-annotator agreement: Cost for increasing n on the TDT3′ topics for both annotators

7. RELATION TO TREC FILTERING

So far we have measured performance using the TDT cost function. We now move on to discuss how our systems (which were trained for the cost metric) perform with respect to the TREC utility metric [21] that was used by the TDT supervised adaptation track [13]. The utility (U) is a linear combination of the number of relevant documents delivered and the number of false alarms:

$$U = W_{Rel} * (\#Targets - \#miss) - \#False\ Alarms$$

W_{Rel} is set equal to 10, weighting relevance 10 times more than non-relevance. To give all topics equal weight the utility is scaled by dividing by $W_{Rel} \times$ *#Targets* to give U_{Norm}

$$U_{Norm} = U/(W_{Rel} \times \#Targets) \quad (1)$$

A minimum utility value (U_{Min}) is considered so as to prevent U_{Norm} from getting skewed by a few bad topics. U_{Min} is typically set at -0.5. This corresponds to an application where the user stops looking at the documents when the non-relevant documents exceeds some threshold. The scaled utility (T11SU) is given as $T11SU = [max(U_{Norm}, U_{Min}) - U_{Min}]/[1 - U_{Min}]$.

T11SU takes a value between 0 and 1. The higher the T11SU, the better the performance. To understand what the values of T11SU mean, note that T11SU and P_{miss} are related as $P_{miss} \leq 1.5(1 - T11SU)$. The equality is reached when there are no false alarms. A system with a T11SU of 0.66 retrieves at least 50% of the relevant documents. A system that declares all documents as NO gets a T11SU of 0.33. With the DP- system we are able to improve the T11SU score from a value of 0.8 for the D- system to a value of 0.86 with no feedback on the TDT3 corpus. One needs to decrease $\#miss$ and $\#fa$ to improve performance on both T11SU and Cost. Yang et al [24] shows how the TDT cost function penalizes the system much more heavily for false alarms and is therefore very recall oriented. Additionally, notice that the TDT cost function penalizes the system for the percentage of false alarms ($P_{FA} = \#fa/\#NonTargets$). Hence, $cost$ gives a system credit for the amount of noise filtered out, since it considers the total number of off-topic stories in the stream (*#NonTargets*). T11SU is stricter and gives no credit for the number of off-topic stories that the system had to process. It only considers the absolute number of false alarms.

T11SU for increasing n is shown in figures 3(a) and 3(b). Systems DP-P and DP-DP again show statistically significant improvements over the D-D system at $n = 20$. System DP-DP is also significantly better than system DP-P. On the TDT4 corpus document feedback without passages (D-D and DP-D) shows fluctuating performance, with utility increasing until $n = 20$ and then dropping off at $n = 2000$. This odd fluctuation can probably be attributed to the trends in relevance that we observed. Adaptation with partially relevant documents that appear later in time probably causes these fluctuations. The T11SU metric is less tolerant to false alarms and hence the effect is probably more pronounced for this metric than for the TDT cost (figures 1). The DP-DP system is consistently better on both metrics.

8. MACHINE TRANSLATION

So far we have only considered English newswire text. As discussed in section 3, the TDT corpus also contains Arabic and Mandarin news stories, which are a significant part of the TDT evaluation. We therefore did the following experiment on the multilingual newswire subsection of the TDT4 corpus. One of the annotators judged the documents as in section 3, but this time the documents also included non-English documents machine translated into English. Machine translation output is very noisy and difficult to read, so we judged only the first 20 documents per topic. Highlighting relevant words, terms and phrases was easy, but reading the text and understanding the content in order to judge the degree of relevance was much harder. 33% of the documents were marked as *don't know* for the document level relevance. The cost and utility are given below:

System	Cost (↓)	Utility (↑)
1. D-D	0.429	0.652
2. P-P	0.283	0.753
3. DP-D	0.297	0.729
4. DP-P	0.208	0.758
5. DP-DP	0.162	0.772

The performance of the five systems is similar to that in section 6. One reason why passage/term level interaction is useful when there are MT documents is that humans can easily detect alternate spellings of named entities. E.g., the title of topic 41002 given to the annotator is the *2001 Nobel Peace Price* with the description mentioning that it was awarded to Kim Dae Jung. Kim Dae Jung (as his name is spelled in the English press) is spelled as *Kim Taicwung* in documents from the Zaobao News Agency (Mandarin translated into English). The *Nobel prize* is spelled as the *Bell prize* in machine translated Zaobao articles. Our annotator could quickly spot these aberrations. Another example is topic 41004, *Murder of the Palestinian Child Mohammed El Dorra. Mohammed* (the spelling in the English sources) is spelled as *Muhammad* in documents translated from Arabic. For a machine to detect these variations we need more sophisticated algorithms [11].

The det plots are given in figure 6.

9. CONCLUSIONS AND FUTURE WORK

The usefulness of document feedback depends on the underlying task and corpus. We have shown that for event tracking, passage

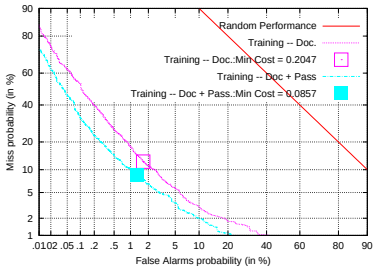


Figure 6: Unsupervised tracking for News-wire in multiple languages.

or term feedback is more effective than document-based feedback. We found that humans can easily give feedback on the relevance of passages and although humans differ in the quantity and portions of text they highlight, performance gains are comparable.

With one training document, it is difficult to ascertain the topic of interest and passage level marking was therefore beneficial. We also showed that the first few relevant documents may be only partially relevant. Thus interaction appears to be a better way to construct a model of the topic. Once the model is fairly good (measured by error rates), we could resort to automatic passage feedback, thus limiting interaction to the early stage of learning.

In this work users were asked to highlight relevant sections of documents. Instead we can show the user a ranked list of passages to judge for relevance in each iteration of interaction. We could also ask for relevance of key-words, named entities etc, in addition or as a substitute for document feedback with the aim of reducing the number of iterations of user-feedback needed.

Acknowledgment

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA), and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

10. REFERENCES

[1] J. Allan. Relevance feedback with too much data. In *SIGIR '95*, pages 337–343, 1995.

[2] J. Allan. *Topic detection and tracking*. Kluwer Academic Publishers, 2002.

[3] J. Allan, V. Lavrenko, and R. Papka. Event tracking. Technical report, UMASS Computer Science Department, 1998.

[4] M. Brown. Abandoning the news. *Carnegie Reporter*, 3(2):2–11, 2005.

[5] J. P. Callan. Passage-level evidence in document retrieval. In *SIGIR '94*, pages 302–310, 1994.

[6] T. Cover and J. Thomas. *Elements of Information theory*. Wiley, New York, 1989.

[7] W. B. Croft and R. Das. Experiments with query acquisition and use in document retrieval systems. In *SIGIR '90*, pages 349–368, 1990.

[8] E. Gabrilovich, S. T. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW*, pages 482–490, 2004.

[9] D. Harman. Relevance feedback revisited. In *SIGIR '92*, pages 1–10, 1992.

[10] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods experiments & analysis, 1998.

[11] K. Knight and J. Graehl. Machine transliteration. In *Proc. of the 8th conf. on European chapter of the ACL*, pages 128–135, 1997.

[12] A. Leuski and J. Allan. Improving realism of topic tracking evaluation. In *SIGIR '02*, pages 89–96, 2002.

[13] Linguistic Data Consortium. *Notebook Proceeding of the TDT 2004 Workshop*, 2004.

[14] J. Makkonen. Investigations on event evolution in tdt. In *HLT-NAACL 2003 Student Workshop*, pages 43–48, 2003.

[15] S. R. Munoz and S. I. Bangdiwala. Interpretation of Kappa and B statistics measures of agreement. *Journal of Applied Statistics*, 24(1):105–112, 1997.

[16] N. N. Ratings. Online newspapers enjoy double-digit year-over-year growth. In `www.netratings.com/pr/pr_051115.pdf`, 2005.

[17] J. Rocchio. *Relevance feedback in information retrieval*. Prentice Hall, 1971.

[18] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *SIGIR 1993*, pages 49–58, 1993.

[19] D. Shaikh. Paper or pixels: What are people reading online? In `http://psychology.wichita.edu/surl/usabilitynews/62/EZprint`, 2004.

[20] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In *Papers of ACM SIGKDD 2000 Workshop on Text Mining*, pages 73–80.

[21] E. M. Voorhees and L. P. Buckland, editors. *TREC 2002)*. Dept of Commerce, NIST, 2002.

[22] H. Xu, B. W. Z. Yang, B. Liu, J. Cheng, Y. L. Z. Yang, and X. Cheng. Trec 11 experiments at CAS-ICT: Filtering and web. 2002.

[23] K. Yang, K. Maglaughlin, L. Meho, and R. G. S. Jr. IRIS at TREC-7. In *TREC-7*, pages 489–500, 1998.

[24] Y. Yang, S. Yoo, J. Zhang, and B. Kisiel. Robustness of adaptive filtering methods in a cross-benchmark evaluation. In *SIGIR '05*, pages 98–105, 2005.