Overview of the TREC 2002 Novelty Track

Donna Harman
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The novelty track was a new track in TREC-11. The basic task was as follows: given a TREC topic and an ordered list of relevant documents (ordered by relevance ranking), find the relevant and "novel" sentences that should be returned to the user from this set. There were 13 groups that participated in this new task.

1 Introduction

The novelty track is a new track in TREC this year and therefore both the results and the evaluation should be viewed as a pilot study. The direct motivation for this track (and the track name) came from Prof. Jaime Carbonnell's talk to the Automatic Summarization Workshop at NAACL in May 2001. In this talk he mentioned that there were other ways to optimize search results than just using relevance ranking alone. One could rank on timeliness of the article, validity of the source of the article, and the comprehensibility and NOVELTY of the information in the article to the user. Whereas the first three of these optimization characteristics are either trivial (timestamps) or very difficult (validity of source and comprehensibility to user), the novelty issue could be operationalized for evaluation by assuming that a user knows nothing at the time of the initial relevant document and all learning happens in the order of document retrieval.

This was the basic design of the novelty track. Given a TREC topic and an ordered list of relevant documents (ordered presumably by relevance ranking alone), find the "novel" information that should be returned to the user from this set. However, unlike earlier work with novelty and redundancy in documents [2], the decision was made to tackle this task at a sentence level. There are several reasons for this decision. First, it was hoped that by reducing the granularity of text unit, it would be easier to identify novel or redundant information. But equally important, the use of sentences combined the novelty task with earlier suggestions for evaluation of passage retrieval. Because the size of a passage is not easily defined (paragraphs are not always available), it was decided to define the "passages" at a sentence level, where the demarcation is clear.

A possible application scenario here would be to envision a smart "next" button that would allow a user to walk down a ranked list of documents by highlighting the next relevant and NOVEL sentence. The user could then view that sentence and if interested, also read the surrounding sentences. Alternatively this task could be viewed as finding key sentences that could be useful as "hot spots" for collecting information to summarize an answer of length X to the information request.

Note that this track is another effort to get beyond the ranked list output for information retrieval. The TREC question-answering track is one approach, but only for direct questions and only for short, fact-based questions. The novelty track explores an alternative approach by returning only relevant AND novel sentences rather than whole documents containing extraneous or duplicate information.

2 Input Data and Task Definition

The basic input data for the novelty track was a set of 50 topics taken from TRECs 6, 7, and 8 (topics 300-450). These 50 topics were selected from the full set of 150 by picking those topics that had between 10 and 70 relevant documents, plus eliminating a few that had large numbers of Federal Register documents, which tend to be very long. This choice was based on having enough relevant documents to work with but not too many for humans to process in creating the truth data. (After NIST staff tried the task, the maximum number of documents was determined to be no more than 25!)

To select the documents once the topics had been picked, NIST modeled the application by selecting actual results from an effective manual run from the appropriate TREC. If there were 25 or fewer relevant documents for the topic, then all the relevant documents were used. If there were more than 25 documents, the top 25 ranked (and relevant) documents from that run were selected for the task. If the run did not find 25 relevant documents, then all the relevant documents it did find were included, along with a random sample of the missing relevant documents up to a maximum of 25 documents. In all cases documents from the Congressional Record (also very long) were NOT selected.

Once the documents were selected, they were ranked at NIST using the ordering produced by same run used in selecting the documents (where possible). Each document was also automatically split into sentences at NIST and sentences were assigned identifiers.

Participants were provided with the topics, the set of sentence-segmented documents, and the order of retrieval for those documents, and were required to process the documents in this order. Each run was to output two ordered sets of sentence identifiers for each of the 50 topics. The first set of sentences was the set of sentences the system determined to contain relevant information. The second set of sentences was the subset of those relevant sentences that the system determined to contain new information, that is, relevant information that was not contained in earlier sentences.

The task was to be done completely automatically. Any fields in the topic could be used, along with any other resources. It was assumed that the set of relevant documents was available as an ordered set, i.e. the entire set could be used in deciding the sentence sets. However the topics had to be processed independently. Both these restrictions reflect the reality of the application.

The content of the topics was a modified version of the original TREC topic statement for each topic, including the original topic fields plus an additional description field (tagged DESC2). This field was what the assessor used as the information need to build the relevant (and novel) sets. Often this was the same as the original description, but many times it was not. The assessors were not supposed to use the narrative field and often the revised description includes some piece of the narrative that they did use.

The test data was released on June 21, 2002, with results due at NIST on September 3.

3 Evaluation

3.1 Creation of truth data

Judgment data was created by having NIST assessors manually perform the task. First they created a file of the relevant sentences and then reduced that file to those that were novel given the document order. Both files were saved. Each topic was independently judged by two different assessors so that the effects of different human opinions could be assessed. Figure 1 presents the exact instructions given to the assessors.

3.2 Analysis of truth data

Since the novelty task was basically requiring systems to automatically select the same sentences that were selected manually by the assessors, it is important to analyze the characteristics of the manually-created truth data in order to better understand the system results. Five particular aspects were examined.

1. what percentage of the sentences in the relevant documents were marked relevant and how does this vary across topics and across assessors?
2. what percentage of the relevant sentences were marked novel and how does this vary across topics and across assessors?
3. how well were the assessors able to pick only "key" sentences as opposed to gathering many consecutive sentences for each relevant piece of information?
4. how different are the results of the two assessors for each topic?
5. where do these differences occur?

Table 1 shows the number of relevant sentences and novel sentences selected for each topic by each of the two assessors that worked on that topic. The column marked "minimum" precedes the results for the assessor who selected the fewest number of relevant sentences. The column marked "rel" is the number of sentences selected as relevant; the next column is the percent of sentences from the total set of relevant documents for that topic that were selected as relevant. The last two columns for each assessor show the number of sentences marked novel and the percentage of the relevant sentences that were marked novel.

Whereas there are large differences in the number of relevant sentences for each topic, there is only a weak topic effect on the percentage of total sentences

Figure 1: Instructions to Assessors

You are trying to create a list of sentences that are:

1. relevant to the question or request made in the description section of the topic,
2. their relevance is independent of any surrounding sentences,
3. they provide new information that has not been found in any previously picked sentences.

Instructions to assessors

1. order printed documents according to the ranked list
2. using the description part of the topic only, go thru each printed document and mark in yellow all sentences that directly provide information requested by the description. Do not mark sentences that are introductory or explanatory in nature. In particular, if there is a set of sentences that provide a single piece of information, only select the sentence that provides the most detail in that set. If two adjacent sentences are needed to provide a single piece of information because of an unusual sentence construction or error in the sentence segmentor, mark both.
3. go to the computer and pull up the online version of your documents. Go through each document, selecting the sentences that you have previously marked (you can change your mind). Save this edited version as "relevant"
4. now go thru the online version looking for duplicate information. Order is important here; if a piece of information has already been picked, then repeats of that same information should be deleted. Instances that give further details of that information should be retained, but instances that summarize details seen earlier should be deleted. Save this second edited version as "new".

per topic selected as relevant. Looking at the results from the "minimum" assessor, the median percentage of sentences marked relevant is about 2%, with the range from 0.2% to 6%. Note that this number is confounded with the assessor effect, as there is a very strong relationship between number of relevant sentences selected and the assessor (see Table 2 for more on this).

It is somewhat surprising that so few of the sentences were selected as relevant. Again using the minimum assessor, 16 of the 50 topics had fewer than 1% of the sentences selected as relevant, whereas only 3 of the 50 topics had more than 5% of the sentences marked relevant.

Equally surprising is the high percentage of relevant sentences that were marked as novel. The median percentage of relevant sentences marked as novel was 93%, with the range from 50% to 100%. In fact, 23 of the 50 topics had ALL relevant sentences marked as novel (by the "minimum" assessor). The fact that the assessors judged most sentences as being novel was a major disappointment of the track. One postulated cause is that the documents being judged were mostly from different sources and different time periods and therefore there was truly little duplicate information. However, it is also possible that judgments for novelty at the sentence level (for long sentences) are likely to always find new information. Next year's novelty track will use multiple sources reporting on the news within the same time period in the hope that a lower percentage of relevant sentences will be marked novel.

Table 2 demonstrates the assessor effect more clearly. The table is ordered by the percentage of the sentences that were marked relevant. The column marked "min" shows the number of topics for which the given assessor had the fewest number of relevant sentences. The next two columns present the total number of relevant sentences found by that assessor and the percentage of sentences in the relevant documents that were marked relevant. The columns marked novel and %rel give the total number of sentences marked novel and the percentage this was of the number marked relevant. The last two columns

Table 1: Analysis of relevant and novel sentences by topic

Topic	minimum	rel	%total	novel	%rel	maximum	rel	%total	novel	%rel
305	A	15	2.01	15	100	B	51	6.84	49	96.08
310	C	0	0	0	0	D	40	7.08	36	90
312	C	5	0.87	5	100	A	18	3.12	12	66.67
314	E	25	2.35	25	100	F	54	5.08	52	96.3
315	C	18	3.08	11	61.11	F	54	9.25	54	100
316	B	22	0.99	18	81.82	G	49	2.2	46	93.88
317	C	23	4.19	23	100	B	33	6.01	22	66.67
322	G	34	4.57	34	100	A	96	12.9	84	87.5
323	G	65	4.57	60	92.31	A	88	6.15	86	97.72
325	G	21	1.25	21	100	F	45	0	lost	0
326	C	10	1	8	80	G	74	7.39	61	82.43
330	E	29	3.49	27	93.1	B	38	4.57	29	76.32
339	C	12	1.6	11	91.67	E	26	3.46	25	96.15
342	E	17	2.17	17	100	G	81	10.32	71	87.65
345	C	47	5.19	47	100	B	65	7.18	52	80
351	C	6	0.75	5	83.33	E	31	3.87	28	90.32
355	F	103	3.94	78	75.73	D	223	8.53	200	89.69
356	D	10	1.2	9	90	B	19	2.29	18	94.74
358	C	40	4.8	37	92.5	E	55	6.6	46	83.64
362	B	47	5.15	46	97.87	A	71	7.79	65	91.55
363	C	11	2.08	10	90.91	A	45	8.52	35	77.78
364	C	42	3.5	42	100	E	45	3.75	45	100
365	G	34	2.8	34	100	E	41	3.38	41	100
368	G	71	4.63	66	92.96	D	121	7.89	94	77.69
369	B	13	1.79	12	92.31	F	30	4.13	25	83.33
377	G	3	0.19	3	100	E	25	1.56	22	88
381	G	19	1.38	19	100	B	38	2.76	29	76.32
382	C	41	1.83	24	58.53	E	114	5.07	110	96.49
384	G	23	1.41	23	100	A	124	7.57	111	89.52
386	A	43	4.3	41	95.35	G	43	4.3	43	100
388	E	56	4.57	56	100	F	123	10.04	96	78.05
394	G	21	2.45	21	100	B	28	3.27	25	89.29
397	C	29	6.18	28	96.5	E	32	6.82	30	93.75
405	B	40	3.75	37	92.5	F	75	7.02	70	93.33
406	G	10	1.79	10	100	C	12	2.14	10	83.33
407	B	32	2.46	29	90.62	A	1301	100	58	4.46
409	B	17	3.26	12	70.59	G	27	5.17	25	92.59
410	E	17	1.92	15	88.24	F	83	9.39	53	63.86
411	C	21	1.86	19	90.48	A	60	5.31	53	88.33
414	E	29	3.62	25	86.21	A	31	3.87	18	58.06
416	E	36	1.96	30	83.33	F	46	2.51	38	82.61
419	A	50	3.32	36	72	D	50	3.32	41	82
420	C	18	1.4	18	100	G	27	2.1	23	85.19
427	E	14	0.31	11	78.57	F	58	1.29	48	82.76
432	E	9	0.96	8	88.89	B	33	3.53	27	81.82
433	C	11	1.72	7	63.64	F	23	3.59	23	100
440	E	19	1.35	19	100	G	107	7.62	98	91.59
445	F	10	1.07	5	50	E	13	1.4	7	53.85
448	C	20	2.25	20	100	D	41	4.62	38	92.68
449	E	57	3.65	57	100	F	81	5.19	71	87.65

Table 2: Analysis of relevant and novel sentences by assessor

Assessor	min	rel	%total	novel	%rel	consecutive	%rel
C	17	348	0.60	343	98.56	97	0.2787
B	6	476	0.82	405	85.08	129	0.2710
D	1	485	0.84	418	86.19	267	0.5505
E	11	690	1.19	644	93.33	193	0.2797
G	10	709	1.23	658	92.81	261	0.3681
F	2	740	1.28	658	88.92	394	0.5324
A	3	1940	3.36	616	31.75	1498	0.7722

present the total number of relevant sentences that were consecutive, and this percentage.

As can be seen from Table 2, there is a major effect from the assessors. For example, assessor "C" was the "minimum" assessor for 17 of the 50 topics; assessor "D" was the minimum assessor for only 1 of the topics. Part of this effect is the normal human variation in relevance judgments as some judges are stricter than others or interpret the question differently. However a second part of the variation comes from different interpretations of the instructions for the task. Note that there is a distinct trend towards more consecutive sentences as there are more relevant sentences selected.

Summarizing the answers to the first three questions:

1. A very low percentage (median 2%) of the sentences in the relevant documents were marked relevant. There was small (0.2% to 6%) but random effect across topics, but definite trends across the assessors in terms of the percentage of sentences (and documents) marked relevant. This is consistent with past findings in the judgment of relevant documents.

2. A very high percentage (median 93%) of the relevant sentences were marked novel. Here there was more variation across topics and less across assessors.

3. In general the assessors were not able to pick only "key" sentences as opposed to gathering consecutive sentences for each relevant piece of information. The percentage of relevant sentences that were consecutive (before or after another relevant sentence) varied from a low of 30% to a high of 77% across assessors.

The final two questions, how different is the output of the two assessors for each topic and where do these differences occur, are explored by Table 3. This table analyzes the differences between the two assessors' relevant sentence judgments. The first two columns give the number of relevant sentences selected by the minimum and maximum assessor for each topic. The next two columns give the sentence coverage and overlap between the two assessors. The overlap is the intersection between their sentences divided by the union of their sentences, i.e. the percentage of the matching sentences divided by sentences that either assessor had selected. The coverage is the intersection of the sentence sets divided by the smaller of those sets, i.e. the percentage of the minimum assessor's sentences that were also chosen by the maximum assessor. The column marked d_cover is the coverage between the documents containing sentences marked relevant, while d_over is the corresponding overlap. The final two columns give the percentage of consecutive sentences selected by the minimum and maximum assessor.

The table is sorted in descending coverage order. Since the minimum assessor has been designated as the "official" assessor in terms of scoring, the coverage metrics are particularly interesting in this evaluation. The average coverage is 0.579, with a median of 0.61. This means that the "second" assessor, in this case the one that picked the larger number of relevant sentences, picked about 60% of the official assessors sentences, plus often many more sentences.

This seemingly low figure is remarkedly similar to that found for relevance judgments for full documents[1], where relevance assessors were shown to agree about 60% of the time that a given document was relevant. But for the novelty task, there are different factors feeding into this disagreement. There is the interpretation of the question and the relative strictness of the judges, factors that are seen in the judgments of full documents. But additionally there is also the issue of how large a section of a given document to select, a factor measured by the number of consecutive sentences.

It was hoped that there would be some clear correlation between some of the measures in Table 3. For example, a low overlap would be correlated with a high number of consecutive sentences. However this is not the case; plots of the coverage or overlap versus the other factors resemble random scatterplots. Illustrations of this can be seen by looking at some of the topics. Topic 427, with a perfect coverage score of 1.00, has an overlap of 0.24. This is due to a large difference in the number of consecutive sentences selected, but ALSO to selection of different sentences as measured by the low document overlap (0.53). Topic 314 has similar number of consecutive sentences selected by each assessor, but the assessors picked different sentences resulting in a low coverage (0.44) and an even lower overlap (0.16).

The complex set of factors governing the differences between assessors makes it unlikely that these differences will lessen. In particular, the assessor instructions to avoid the use of consecutive sentences did not work, and it is unlikely that agreement would have improved even if there had been fewer consecutive sentences. In the next running of the novelty track this instruction will be removed.

What remains to be done given this low level of agreement is to understand how the system comparisons are effected by all this noise. Voorhees [1] showed that there was no effect as long as enough topics were used for averaging. Part of the analysis next year will be to check if this also holds true for sentence relevance in the novelty track.

3.3 Scoring

The sentences selected manually by the NIST assessors were considered the truth data. Obviously we could have chosen one set of assessors as the official one (similar to TREC ad hoc), and use the second set only for human agreement measurements. However the decision was made to use the truth data from the assessor that marked the smallest number of relevant sentences (on a per topic basis) as the main score. The reason for this is that the biggest disagreement between assessors had to do with how many sequential sentences they took as relevant. Often they included sentences "for context", even though the instructions tried to discourage this. By taking the minimum of two assessors, it was hoped to avoid many of the disagreements. This definition also matches better with the stated goals of the track. Note that one assessor disagreed with the original assessor's relevance judgments for topic 310 and could find no relevant sentences in any of the documents. We eliminated that topic from the final test set, so scores were computed over the remaining 49 topics.

Participants were told that the scoring would be based on the smaller set before runs were submitted, but, of course, they did not have access to the assessor sentence sets.

The track guidelines specified sentence set recall and precision as the evaluation measures for the track. Let M be the number of matched sentences, i.e., the number of sentences selected by both the assessor and the system, A be the number of sentences selected by the assessor, and S be the number of sentences selected by the system. Then sentence set recall is M/A and precision is M/S.

As previous filtering tracks have demonstrated, set-based recall and precision do not average well, especially when the assessor set sizes vary widely across topics. Consider the following example as an illustration of the problems. One topic has hundreds of relevant sentences and the system retrieves 1 relevant sentence. The second topic has 1 relevant sentence and the system retrieves hundreds of sentences. The average for both recall and precision over these two topics is approximately .5 (the scores on the first topic are 1.0 for precision and essentially 0.0 for recall, and the scores for the second topic are the reverse), even though the system did precisely the wrong thing. While most real submissions won't exhibit this extreme behavior, the fact remains that recall and precision averaged over a set of topics is not a good diagnostic indicator of system performance. There is also the problem of how to define precision when the system returns no sentences ($S = 0$). Not counting that question in the evaluation for that run means different systems are evaluated over different numbers of topics, while defining precision to be either 1 or 0 is extreme. (The average scores given in Appendix A defined precision to be 0 when $S = 0$ since that seems the least evil choice.)

To avoid these problems, the primary measure reported for novelty track runs is the F measure, defined as

$$\mathrm{F} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

The average of the F measure is meaningful even when the judgment sets sizes vary widely. For example, the F measure in the scenario above is essentially 0, an intuitively appropriate score for such behavior. Using the F measure also deals with the problem of what to do when the system returns no sentences since recall is 0 and the F measure is legitimately 0 regardless of what precision is defined to be.

Table 4: Organizations participating in the TREC 2002 Novelty Track

Carnegie Mellon University
Columbia University
Fudan University
IRIT/SIG
National Taiwan University
NTT Communication Science Labs
Queens College, CUNY
Streamsage
Tsinghua University
University of Amsterdam/ILLC
University of Iowa
University of Massachusetts at Amherst
University of Michigan

4 Participants and Descriptions of Approaches

Table 4 lists the 13 groups that participated in the TREC 2002 novelty track. The rest of this section contains short summaries submitted by most of the groups about their approaches to the novelty task.

4.1 Carnegie Mellon University

To find relevant sentences we used a simple baseline of cosine similarity with tf.idf weighting and pseudo-relevance feedback, treating sentences as very short documents. We tried different classifiers using lexical and semantic features derived from a simple parse as well as proximity to sentences with very high tf.idf scores. To model redundant sentences we used a very simple translation model. The translation probabilities for word or short phrase pairs were based on the skew divergence between word distributions derived from a mixture model of unigrams extracted from WordNet relations. The parse tree for each sentence was transformed into a graph of modifier relations. The overall redundancy measure between sentences was then calculated using a basic greedy graph-matching algorithm.

4.2 Columbia University

Our principal interest in the Novelty Track was to experiment with ideas we are developing in the detection of new information, but we found that the relevance part of the task here absorbed most of the time that we had alloted and most of our attention. We decided that it would be more interesting to adapt our new information tools to the relevance part than to try to use established IR strategies. Our approach on relevance question was to expand the information in all fields of the given topics with 1) semantic equivalents of the content words in the topic, and 2) related words determined by co-occurrence statistics from a background corpus. Sentences were selected largely by the number of words that match the expanded queries. In addition, we reclustered the set of relevant documents and in some cases eliminated many documents from inclusion. We had little time left for the novelty part and relied solely on the overlap of the topic words and their semantic equivalents.

4.3 IRIT/SIG

IRIT developed a new strategy in order to detect the relevant sentence that we did not try in a more general context of document retrieval but did try previously in document categorization. In our approach a sentence is considered as relevant if it matches the topic with a certain level of coverage. This level of coverage depends on the category of the words. Three types of words have been defined: highly relevant, lowly relevant and no relevant. With regard to the novelty part, a sentence is considered as novel with regard to a topic when its level of coverage with the previously processed sentences and with the best-retrieved sentences does not exceed a certain threshold.

4.4 National Taiwan University

In the novelty task, the amount of information that can be used in a sentence is the major challenging issue. Some sort of information expansion method was introduced to tackle this problem. Our approach to relevance identification was to expand the information of a sentence with the context of this sentence using a sliding window method. The similarity was measured by the number of words of a topic description that match the sentences within a window. Besides, WordNet was employed to relax word match operation to inexact match. In the novelty detection part, we first applied a coherent text segmentation algorithm to partition the sentences extracted from the relevance identification part into several coherent passages denoting sub-topics. Then we compute the similarity of each sentence with each passage. A sentence was in terms of a sentence-passage similarity vector. Two sentences are regarded as similar if they are related to the same sub-topics . In this way, the redundant sentences were filtered out.

4.5 NTT Communication Science Labs

Our approach is based on query-biased multi-document summarization methods. "Relevant" sentences are selected from each document using Support Vector Machines(SVMs), which are trained on a query-sentence data set. The data set consists of query-sentence pairs whose relevance are judged by us and the queries are chosen from the TREC topics that are not used in the novelty track. "New" sentences are chosen from the relevant sentences based on Maximum Marginal Relevance (MMR) measure.

4.6 Queens College, CUNY

For this experiment, we employ all sections of a topic to form long queries for retrieval because the "documents" are actually sentences. The queries average to 938/49 unique terms. Since the sentences come from relevant documents of TREC-8, we use the TREC-8 dictionary to provide better statistics for processing and retrieval. However, the high Zipf threshold has been reset to 400,000 to include more high frequency terms.

Only initial retrieval without pseudo-relevance feedback was performed. Based on experimentation with the four training topics, we decide to test two RSV threshold (tr) values to help decide on relevance of retrieved sentences: submission pircs2N01/2 use tr=1.25, and pircs2N03/4 use tr=1.5.

This set of relevant sentences is sorted according to DocID. For each sentence, every one of its unstemmed words is expanded with synonyms by consulting with WordNet. The resultant set of words is sorted and duplicates removed. A double loop passes down the sentence list, and a novelty coefficient based on the Dice formula is evaluated for each pair of sentences S_i and S_j. The novelty coefficient is defined as the intersection of S_i and S_j divided by the union of S_i and S_j.

If the novelty coefficient is less than a threshold tv, S_j is considered novel with respect to S_i, otherwise S_j is removed. pircs2N01 and pircs2N03 employ a threshold tv=0.3, and pircs2N02, pircs2N04 use tv=0.5. In addition, a fifth submitted run pircs2N05 does not use synonyms, just raw words, and acts as control. Its thresholds are tr=1.5, tv=0.3.

4.7 Streamsage

I concentrated on query expansion, using our New York Times news story corpus, and the description field. Starting with the nouns in the topic title, I searched our corpus for multi-word units containing them (subject to grammatical and frequency constraints), then added other nouns in the MWU as search terms. I also included nouns from desc in the same noun phrase as a title noun. Sentences which contained a search term were returned as relevant.

4.8 Tsinghua University

In this year's novelty track, we performed two-step research to find relevant sentence, and then to eliminate repetitive information. On finding relevant information, our work focused on four parts:

1. Extracting key information in topics. We classified words in the topic into three classes by statistical learning and rule-based learning: useful keywords, general describing words and negative words.

2. Finding efficient query expansion (QE) techniques. Besides thesaurus based QE, we proposed and studied a new statistical expansion approach, which expands terms that co-occurred in a fixed window size with title words in the relevant document set, called local co-occurrence expansion. The results are extremely good.

3. document/sentence term expansion (DE). Sometimes, the query mentions a general topic while some relevant documents describe detailed information. In this case, QE may not help because you do not know to what extent the terms should be expanded. We proposed term expansion in documents (referred as DE) to solve the problem.

4. topic classification. QE and DE are oriented from two aspects of retrieval problem and may work well for different types of topics. Therefore we classified the topics into two classes according to similarities between topic fields to perform QE or DE respectively, which leads to better performance than either approach. On eliminating repetitive information, rather than concept of similarity, we used the concept of unsymmetrical sentence overlapping. It represents the extent of the information taken by one sentence overlapped by another one. Our experimental results show it is better than the symmetrical measure

of similarity. Two different elimination strategies are studied. One is sentence to sentence comparison, the other one is sentence to pool overlapping technique. In our experiments, the performances of two strategies are almost equal.

4.9 University of Amsterdam/ILLC

For identifying *relevant* sentences, we used a fairly minimal approach. For a given topic, the sentences in the relevant documents for that topic were viewed as documents themselves, and we ran the topic against this sentences-as-documents collection using a retrieval engine based on a standard vector space model, with the tfv.nfx weighting scheme. Three different runs were submitted: one where all documents were stemmed, a second where they were lemmatized, and in the third run the results of the other two runs were simply merged.

For identifying the list of *new* sentences, we scanned the list of relevant sentences and filtered out sentences that were entailed by the sentences kept so far. Our notion of entailment between two text segments s_1 and s_2 is based on comparing the sum of the inverted document frequencies of the terms that occur in both s_1 and s_2 to the inverted document frequencies of the terms occurring in s_2. If it is beyond a certain threshold, this entailment score prevents a sentence s_2 from being added to the set of new sentences s_1 that we have built up so far.

4.10 University of Massachusetts at Amherst

Our approach was to apply standard techniques that have proven successful for document retrieval and filtering to see if they also work well at the sentence level. We began by building a larger training corpus from 48 of the TREC ad-hoc retrieval track topics, supplementing the handful of training topics that NIST provided. Our methods for building this corpus followed the general specifications for building the test collection. We used the same instructions provided to the NIST assessors, though we used undergraduates rather than retired analysts to do the assessments.

The task was then treated as two separate problems: (1) identify the relevant sentences then (2) filter out the redundant sentences. In identifying relevant sentences, we experimented with several retrieval models, including language modeling and the vector space model with tf-idf weighting. We found that the tf-idf approach worked best on our training data, so both of our systems (CIIR02tfkl and CIIR02tfnew) use that method to identify relevant sentences.

For novelty filtering, we built two different systems. One (CIIR02tfkl) uses the Kullback-Leibler divergence between a sentence and all previously seen sentences to assign novelty scores. The other (CIIR02tfnew) employs a set-difference approach by counting the number of previously unseen words in each relevant sentence. On both the training and the test data, the set difference system outperformed the language modeling system when applied to our own relevance results. However, for both collections, the language modeling approach performed better when applied to the known relevant sentences–i.e., if we "cheat" and use the truth data as a preliminary step.

4.11 University of Michigan

The Michigan novelty detection systems for this year's evaluation were based on the MEAD multi-document, extractive summarizer. Using the sample data, we first concentrated on modifying MEAD to better detect sentences that are relevant to a user's query. In particular, our modifications tried to capture our observation that the humans who judged the sample clusters tended to choose groups of relevant sentences, rather than individual sentences from different places in a source document. To do this, we implemented a new sentence reranker within MEAD, which favored groups of sentences with relatively high concentrations of key words relevant to the overall cluster of documents. We also developed some new sentence features within MEAD, which measured how related a given sentence is to a user's query. Specifically, we used a query-title-word-overlap feature, which quanified the extent to which a given sentence contained words that were present in the title of the user query.

Our aim main in participating in the novelty track was to set a simple base line for future, linguistically motivated experiments on the task.

5 Results

Thirteen groups submitted 43 runs to the novelty track. For all runs, the F score for the relevant sentence sets was greater than the score for the new sentence sets. This suggests that finding the relevant parts of a document is somewhat easier than finding the nonredundant parts.

Since the novelty track was a completely new task, groups had no training data and little idea of what to expect. One group (the University of Massachusetts

Table 5: Average F scores for baseline and system results for the Novelty track.

	Relevant	New
Second human judges	0.371	0.353
Random sentences	0.040	0.036
thunv1	0.235	0.217
thunv2	0.235	0.216
thunv3	0.235	0.216
CIIR02tfnew	0.211	0.209
thunv4	0.225	0.206
CIIR02tfkl	0.211	0.196
pircs2N02	0.209	0.193
pircs2N01	0.209	0.188
pircs2N04	0.197	0.184
ss1	0.186	0.183

at Amherst) constructed extensive training data following the assessor instructions and used this to guide their research (run tags CIIR). Another high scoring group, City University of New York (run tags pircs), used traditional information retrieval methods, treating the sentences as documents. Tsinghua University (run tags thunv) used a completely new method devised especially for this task.

One of the requirements for a new track is to do sanity-checking of the evaluation itself. To this end, NIST computed the average F score for the second human assessor sentence sets and for sets of sentences randomly selected from the target documents. The results are shown in Table 5, which also includes the scores for some of the best runs for comparison. The scores for the systems fall in between the human and random performance, support for a claim that the evaluation is credible.

The track will be run again in 2003, with topics specifically constructed for the task. The data will consist of several news sources from the same time period in hopes that there will be more duplicate information.

References

[1] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[2] Yi Zhang, Jamie Callan, and Thomas Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88, 2002.

Table 3: Analysis of sentence coverage and overlap, document coverage and overlap and percent of consecutive sentences by topic

Topic	min.rel	max.rel	coverage	overlap	d_coverage	d_overlap	%min consecutive	%max consecutive
427	14	58	1.00	0.24	1.00	0.53	7.14	41.38
407	32	1301	1.00	0.02	1.00	0.52	34.38	98.08
397	28	32	0.89	0.71	1.00	1.00	7.14	15.62
315	18	54	0.89	0.29	1.00	0.40	11.11	27.78
305	15	51	0.87	0.25	0.83	0.62	40	58.82
365	34	41	0.82	0.60	0.81	0.65	29.41	24.39
414	29	31	0.79	0.62	0.95	0.83	17.24	9.68
440	19	107	0.79	0.14	0.91	0.42	10.53	48.6
364	42	45	0.74	0.55	0.90	0.82	30.95	42.22
433	11	23	0.73	0.31	0.83	0.45	9.09	43.48
355	103	223	0.73	0.30	0.94	0.62	54.37	69.51
322	34	96	0.71	0.23	0.75	0.52	23.53	43.75
419	50	50	0.70	0.54	1.00	0.88	30	22
358	40	55	0.68	0.40	1.00	0.72	22.5	16.36
368	71	121	0.68	0.33	0.96	0.88	39.44	61.16
432	9	33	0.67	0.17	0.88	0.39	11.11	27.27
382	24	114	0.67	0.13	1.00	0.41	20.83	37.72
377	3	25	0.67	0.08	0.67	0.29	0	52
362	47	71	0.66	0.36	1.00	1.00	17.02	15.49
405	40	75	0.65	0.29	0.94	0.75	22.5	50.67
448	20	41	0.65	0.27	0.80	0.50	15	29.27
388	56	123	0.64	0.25	0.78	0.64	39.29	73.17
363	11	45	0.64	0.14	0.70	0.44	9.09	22.22
330	29	38	0.62	0.37	1.00	0.87	17.24	21.05
384	23	124	0.61	0.11	0.89	0.36	21.74	50.81
312	5	18	0.60	0.15	1.00	0.10	40	0
326	10	74	0.60	0.08	1.00	0.20	30	45.95
410	17	83	0.59	0.11	0.93	0.62	5.88	45.78
339	12	26	0.58	0.23	1.00	0.67	8.33	42.31
323	65	86	0.55	0.31	0.94	0.73	46.15	50
369	13	30	0.54	0.19	1.00	0.80	15.38	43.33
342	17	81	0.53	0.10	0.92	0.67	11.76	60.49
406	10	12	0.50	0.29	0.83	0.56	0	33.33
356	10	19	0.50	0.21	0.83	0.42	10	15.79
351	6	31	0.50	0.09	0.83	0.33	0	25.81
394	21	28	0.48	0.26	0.83	0.62	28.57	25
317	23	33	0.48	0.24	1.00	0.70	47.83	36.36
409	17	27	0.47	0.22	0.85	0.52	11.76	0
345	47	65	0.45	0.23	1.00	0.64	38.3	21.54
314	25	54	0.44	0.16	1.00	0.32	44	42.59
386	43	43	0.40	0.25	1.00	0.88	32.56	20.93
416	36	46	0.39	0.21	0.83	0.48	11.11	39.13
411	21	60	0.38	0.11	0.86	0.30	33.33	25
449	57	81	0.37	0.18	0.86	0.40	35.09	80.25
445	10	13	0.30	0.15	1.00	0.40	40	7.69
381	19	38	0.26	0.10	0.80	0.63	0	21.05
325	21	45	0.00	0.00	1.00	0.47	9.52	0
420	18	27	0.00	0.00	1.00	0.05	83.33	18.52
310	0	40	0.00	0.00	0.00	0.00	35	0