Overview of the TREC 2003 Novelty Track

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

Abstract

The novelty track was first introduced in TREC 2002. Given a TREC topic and an ordered list of documents, systems must find the relevant and novel sentences that should be returned to the user from this set. This task integrates aspects of passage retrieval and information filtering. This year, rather than using old TREC topics and documents, we developed fifty new topics specifically for the novelty track. These topics were of two classes: "events" and "opinions". Additionally, the documents were ordered chronologically, rather than according to a retrieval status value. There were four tasks which provided systems with varying amounts of relevance or novelty information as training data. Fourteen groups participated in the track this year.

1 Introduction

The novelty track was introduced as a new track last year [5]. The basic task is as follows: given a topic and an ordered set of relevant documents segmented into sentences, return sentences that are both relevant to the topic and novel given what has already been seen. This task models an application where a user is skimming a set of documents, and the system highlights new, on-topic information.

There are two problems that participants must solve in the novelty track. The first is identifying relevant sentences, which is essentially a passage retrieval task. Sentence retrieval differs from document retrieval because there is much less text to work with, and identifying a relevant sentence may involve examining the sentence in the context of those surrounding it. We have specified the unit of retrieval as the sentence in order to standardize the task across a variety of passage retrieval approaches, as well as to simplify the evaluation.

The second problem is that of identifying those relevant sentences that contain new information. The operational definition of "new" is information that has not appeared previously in this topic's set of documents. In other words, we allow the system to assume that the user is most concerned about finding new information in this particular set of documents, and is tolerant of reading information he already knows because of his background knowledge. Since each sentence adds to the user's knowledge, and later sentences are to be retrieved only if they contain new information, novelty retrieval resembles a filtering task.

To allow participants to focus on the filtering and passage retrieval aspects separately, this year the track offered four tasks. The base task was to identify all relevant and novel sentences in the documents. The other tasks provided varying amounts of relevant and novel sentences as training data. Some groups which chose to focus on passage retrieval alone did only relevant sentence retrieval in the first task.

2 Input Data

Last year, the track used 50 topics from TRECs 6, 7, and 8, along with relevant documents in rank order according to a top-performing manual TREC run. The assessors' judgments for those topics were remarkable in that almost no sentences were judged to be relevant, despite the documents themselves being relevant. As a consequence, nearly every relevant sentence was novel. This was due in large part to assessor disagreement (the assessors were not the original topic authors) and drift (the document judgments were all made several years ago).

To both solve the assessor drift problem and to achieve greater redundancy in the test data, this year we constructed fifty new topics on a collection of three contemporaneous newswires. For each topic, the assessor composed the topic, selected 25 relevant documents by searching the collection, and labeled the relevant and novel sentences in the documents.

As an added twist, 28 of the topics concerned events such as the bombing at the 1996 Olympics in Atlanta, while the remaining topics focused on opinions about controversial subjects such as cloning, gun control, and same-sex marriages. The topic type was indicated in the topic description by a `<toptype>` tag.

The documents for the novelty track were taken from the AQUAINT collection. This collection is unique in that it contains three news sources from overlapping time periods: New York Times News Service (Jun 1998 – Sep 2000), AP (also Jun 1998 – Sep 2000), and Xinhua News Service (Jan 1996 – Sep 2000). We intended that this collection would exhibit greater redundancy and thus less novel information, increasing the realism of the task. The assessors, in creating their topics, searched the AQUAINT collection using WebPRISE, NIST's IR system, and collected 25 documents which they deemed to be relevant to the topic.

Once selected, the documents were ordered chronologically. (Chronological ordering is achieved trivially in the AQUAINT collection by sorting document IDs.) This is a significant change from last year's task, in which they were ordered according to retrieval status value in a particular TREC ad hoc run. Last year's ordering was motivated by the idea of seeking novel information in a ranked list of documents, whereas this year, the task more closely resembles reading new documents over time. This approach seems to make more sense when working with news articles, since background information tends to occur more completely in earlier articles and is summarized more briefly as time goes on and new information is reported. With relevance ranking, one can identify novel sentences but there is no sense of which document should come first.

The documents were then split into sentences, each sentence receiving an identifier, and all sentences were concatenated together to produce the document set for a topic.

3 Task Definition

This year, there were four tasks:

Task 1. Given the set of 25 relevant documents for the topic, identify all relevant and novel sentences. (This was the same as last year's task.)

Task 2. Given the relevant sentences in all 25 documents, identify all novel sentences.

Task 3. Given the relevant and novel sentences in the first 5 documents **only**, find the relevant and novel sentences in the remaining 20 documents.

Task 4. Given the relevant sentences from all 25 documents, and the novel sentences from the first 5 documents, find the novel sentences in the last 20 documents.

These four tasks allowed the participants to test their approaches to novelty detection given different levels of training: none, partial, or complete relevance information, and none or partial novelty information.

Participants were provided with the topics, the set of sentence-segmented documents, and the chronological order for those documents. For tasks 2-4, training data in the form of relevant and novel "sentence qrels" were also given. The data were released and results were submitted in stages to limit "leakage" of training data between tasks. Depending on the task, the system was to output the identifiers of sentences which the system determined to contain relevant and/or novel relevant information.

4 Evaluation

4.1 Creation of truth data

Judgments were created by having NIST assessors manually perform the task. From the concatenated document set, the assessor selected the relevant sentences, then selected those relevant sentences that were novel. Each topic was independently judged by two different assessors, the topic author and a "secondary" assessor, so that the effects of different human opinions could be assessed.

4.2 Analysis of truth data

Since the novelty task requires 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. In particular, there were several concerns raised by the peculiarities of last year's data.

1. What percentage of the sentences were marked relevant, and how does this vary across topics and across assessors?

2. Did the quantity of relevant and new information improve from last year? In particular, are more sentences relevant, and are fewer relevant sentences novel?

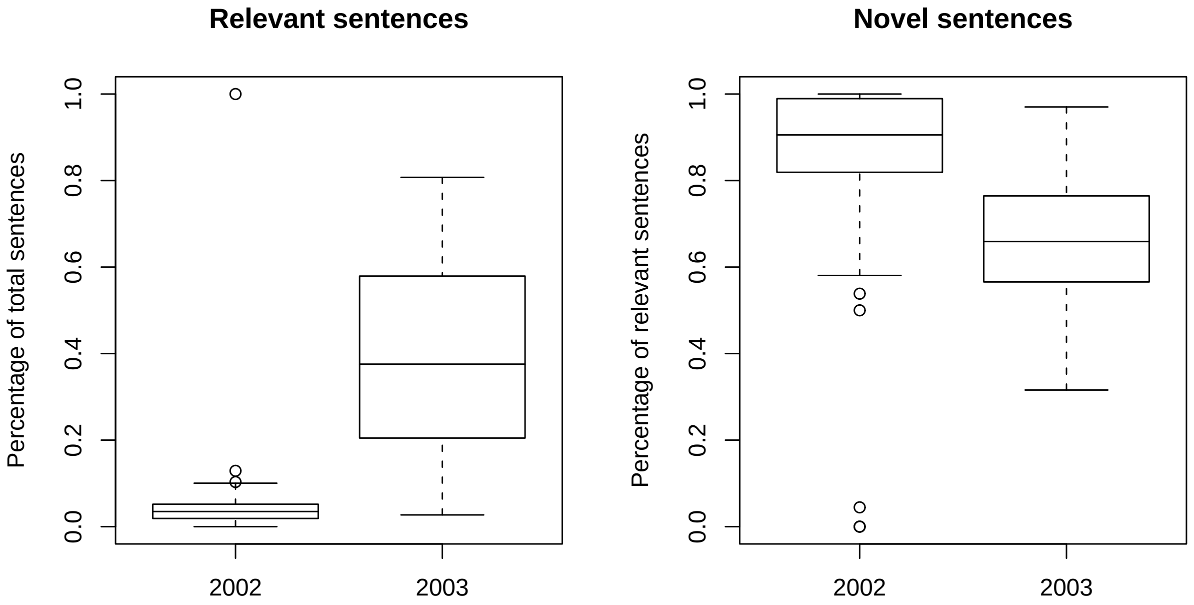


Figure 1: Percentage of relevant and novel sentences (both primary and secondary assessors), compared to 2002 (both minimum and maximum assessors).

3. How different are the results of the secondary assessor from the primary assessor who authored the topic and selected the documents?

4. Is there any difference between "event topics" and "opinion topics", in terms of amounts of relevant and new information?

Table 1 shows the number of relevant and novel sentences selected for each topic by each of the two assessors who worked on that topic. The column marked "assr-1" precedes the results for the primary assessor, whereas "assr-2" precedes those of the secondary assessor. The column marked "rel" is the number of sentences selected as relevant; the next column, "%total", is the percentage of the total set of sentences for that topic that were selected as relevant. The column marked "new" gives the number of sentences selected as novel; the next column, "%rel", is the percentage of relevant sentences that were marked novel. The column "sents" gives the total number of sentences for that topic, and "type" indicates whether the topic is about an event (**E**) or about opinions on a subject (**O**).

One of the most striking aspects of Table 1 is the difference in relevant and new percentages from last year. The median percentage of relevant sentences is 37.56%, compared with about 2% last year. For novel sentences, the median is 65.91%, compared with 93% last year. Figure 1 illustrates the range of relevant and novel sentences, and compares it to the 2002 data. Whereas last year, almost no sentences were selected as relevant, and as a result nearly every relevant sentence was novel, this year the distributions of relevant and novel sentences are much more reasonable.

The analysis of assessor effects is complicated by the fact that only four of the seven assessors (B, C, D, and E) acted as both primary and secondary assessors. Assessor A only judged as a primary assessor, and assessors F and G only judged as secondary assessors (i.e., they judged other assessors topics, but did not author their own).

As we might expect, there is a large effect from the assessors. For relevant sentence selection, this effect is more significant than either topic type or judgment round. The four assessors who judged topics in both rounds (B, C, D, and E) were quite different from each other, but judged similarly from the first round to the second. For novel sentences, it's a different story; differences between assessors are more pronounced in the first round, but in the second they are all quite similar to each other. Overall, the number of novel sentences selected is more uniform across

Table 1: Analysis of relevant and novel sentences by topic

Topic	type	sents	assr-1	rel	%total	new	%rel	assr-2	rel	%total	new	%rel
N1	O	880	A	184	20.91	151	82.07	F	457	51.93	265	57.99
N2	E	500	D	78	15.6	43	55.13	B	170	34.0	58	34.12
N3	E	932	C	596	63.95	331	55.54	E	248	26.61	152	61.29
N4	E	928	B	438	47.2	265	60.5	D	113	12.18	72	63.72
N5	E	1662	B	259	15.58	219	84.56	G	293	17.63	246	83.96
N6	E	424	B	317	74.76	233	73.5	C	294	69.34	192	65.31
N7	E	306	E	95	31.05	79	83.16	D	99	32.35	61	61.62
N8	E	659	D	158	23.98	107	67.72	G	349	52.96	210	60.17
N9	E	637	B	160	25.12	62	38.75	F	263	41.29	205	77.95
N10	E	257	C	191	74.32	159	83.25	F	174	67.7	139	79.89
N11	E	393	C	148	37.66	108	72.97	G	130	33.08	107	82.31
N12	O	1044	C	729	69.83	579	79.42	D	76	7.28	62	81.58
N13	O	941	A	205	21.79	166	80.98	F	457	48.57	198	43.33
N14	O	1129	D	93	8.24	42	45.16	G	191	16.92	122	63.87
N15	O	649	C	522	80.43	421	80.65	B	378	58.24	214	56.61
N16	E	500	E	179	35.8	119	66.48	F	274	54.8	183	66.79
N17	O	1106	B	792	71.61	488	61.62	G	724	65.46	524	72.38
N18	O	1238	B	537	43.38	429	79.89	D	98	7.92	47	47.96
N19	O	867	D	62	7.15	37	59.68	C	423	48.79	253	59.81
N20	O	886	D	69	7.79	41	59.42	B	228	25.73	169	74.12
N21	O	932	E	340	36.48	194	57.06	G	317	34.01	265	83.6
N22	O	841	D	84	9.99	52	61.9	F	401	47.68	295	73.57
N23	O	896	E	346	38.62	254	73.41	D	134	14.96	86	64.18
N24	O	968	E	160	16.53	76	47.5	B	220	22.73	101	45.91
N25	O	701	D	19	2.71	16	84.21	C	301	42.94	249	82.72
N26	O	911	C	661	72.56	283	42.81	E	178	19.54	137	76.97
N27	O	962	C	730	75.88	577	79.04	E	273	28.38	229	83.88
N28	O	978	B	371	37.93	261	70.35	E	109	11.15	80	73.39
N29	O	861	D	65	7.55	52	80.0	B	69	8.01	39	56.52
N30	O	900	C	445	49.44	307	68.99	G	497	55.22	386	77.67
N31	O	1220	C	985	80.74	752	76.35	D	74	6.07	48	64.86
N32	O	1078	B	216	20.04	100	46.3	C	684	63.45	475	69.44
N33	E	680	C	526	77.35	376	71.48	G	441	64.85	297	67.35
N34	E	1030	E	475	46.12	217	45.68	D	106	10.29	78	73.58
N35	E	399	E	221	55.39	77	34.84	B	253	63.41	95	37.55
N36	E	355	C	167	47.04	162	97.01	F	239	67.32	183	76.57
N37	E	547	D	76	13.89	52	68.42	G	263	48.08	196	74.52
N38	O	1127	D	140	12.42	96	68.57	F	252	22.36	188	74.6
N39	E	590	B	211	35.76	128	60.66	E	221	37.46	151	68.33
N40	E	533	B	307	57.6	183	59.61	E	209	39.21	145	69.38
N41	E	672	C	535	79.61	403	75.33	B	297	44.2	99	33.33
N42	E	1119	C	400	35.75	340	85.0	B	414	37.0	140	33.82
N43	E	224	E	122	54.46	67	54.92	C	142	63.39	116	81.69
N44	E	585	C	432	73.85	266	61.57	D	55	9.4	34	61.82
N45	E	1054	E	262	24.86	124	47.33	F	412	39.09	254	61.65
N46	E	549	E	322	58.65	123	38.2	C	342	62.3	165	48.25
N47	E	601	C	420	69.88	290	69.05	B	142	23.63	86	60.56
N48	E	1075	D	98	9.12	53	54.08	E	245	22.79	156	63.67
N49	E	684	D	209	30.56	66	31.58	C	225	32.89	147	65.33
N50	E	810	E	400	49.38	200	50.0	C	485	59.88	182	37.53

assessors than relevant sentences. Figure 2 illustrates these differences.

Last year, we found that the assessors tended to pick consecutive groups of sentences as relevant, despite being instructed otherwise. This year, we did not restrict them from selecting consecutive sentences, instead allowing them to select whatever they felt was necessary. As might be expected, this along with the greater amount of relevant sentences chosen resulted in a much higher occurrence of consecutive relevant sentences. On average, 84% of relevant sentences were selected immediately adjacent to another relevant sentence. The median length of a string of consecutive relevant sentences was 2; the mean was 4.252 sentences.

Overall, there was not a large difference between the primary and secondary assessor in terms of the number of relevant and novel sentences selected. Figure 3(a) shows that the secondary assessors tended to be a little more restrictive in their judgments, but this difference is not statistically significant. This implies that the marked difference in judgment patterns we see between this year and last is not only due to an assessor effect. Having more recent documents and topics, and allowing the assessors to select the relevant documents, probably also played a role.

There is a larger difference between event and opinion topics. Figure 3(b) illustrates this. Opinion topics tended to have a lower percentage of relevant and a higher percentage of novel sentences than events. The higher percentage of novel sentences is actually due to the lower percentage of relevant sentences. The difference is statistically significant for relevant sentences, but not for novel ones.

While it may be the case that having multiple news sources from the same time period increased redundancy over last year's topics, having stories from two or three wires did not make a significant difference in the number of novel sentences. Only one topic (10) drew stories from a single news source; all others involved either two or three sources. On average, 63.61% of relevant sentences were novel for topics with two sources, and 64.73% for those with three. Both of these are less than the new percentage for topic 10 (83.25%), but with only one topic we can't make any conclusions.

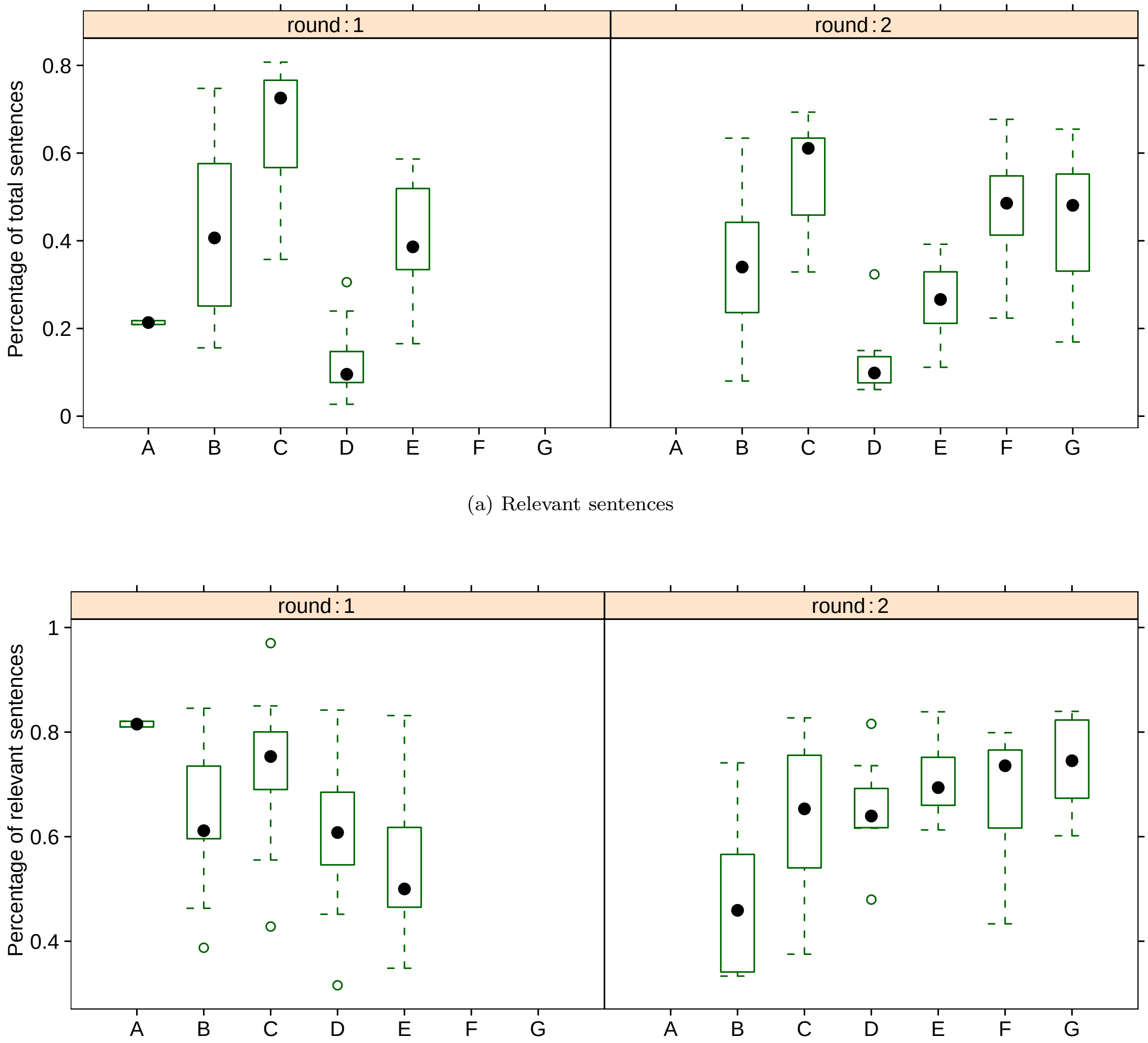
To summarize, the topics and judgments are much improved over last year. While there are differences in judging between the two assessment rounds, and between the different topic types, once again differences between assessors are dominant. Differences are more marked for relevant sentence selection than for novelty, indicating that there is a real difference between these two tasks.

4.3 Scoring

The sentences selected manually by the NIST assessors were considered the truth data. In contrast to last year, where concerns about assessors selecting groups of sentences for context drove the evaluation to use the assessor with the fewest selected relevant sentences (the so-called "minimum assessor"), this year the judgments by the topic author were taken as the truth data. The judgments by the secondary assessor were taken as a human baseline performance in the task.

Because relevant and novel sentences are returned as an unranked set in the novelty track, we cannot use traditional measures of ranked retrieval effectiveness such as mean average precision. The track guidelines specified the F measure as the primary evaluation measure for the track. The F measure (from van Rijsbergen's E measure) is itself derived from set precision and recall. For the novelty track, the "set" in question is the set of retrieved sentences (rather than documents as in the retrieval case). Relevant and novel sentence retrieval are evaluated separately. 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.

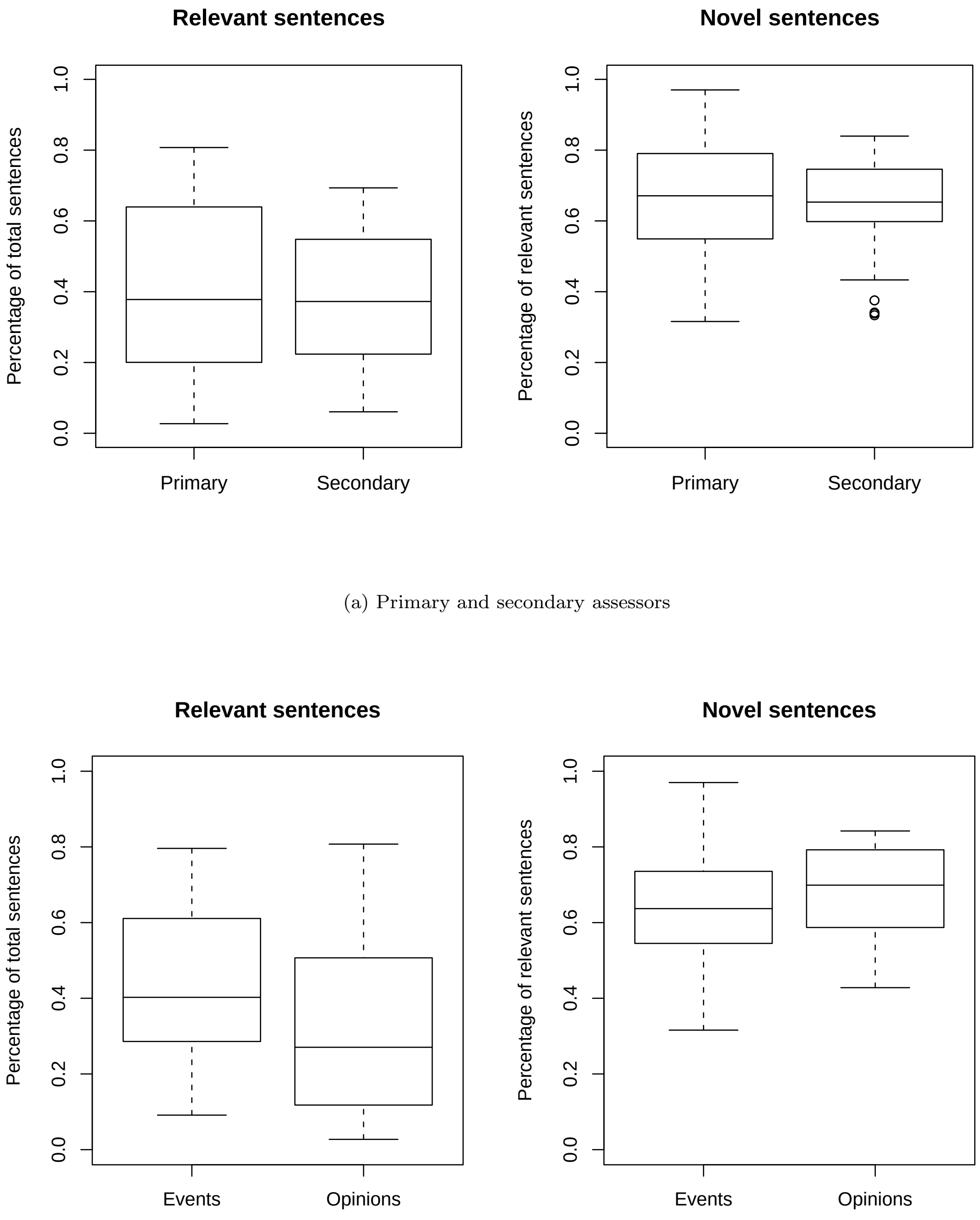
As previous filtering tracks have demonstrated, set-based recall and precision do not average well, especially when the assessor set sizes vary widely across topics. Consider the following example as an illustration of the problems. One topic has hundreds of relevant sentences and the system retrieves 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 differ-



(a) Relevant sentences

(b) New sentences

Figure 2: Assessor effects.



(a) Primary and secondary assessors

(b) Event and opinion topics

Figure 3: Differences between assessment rounds and topic types.

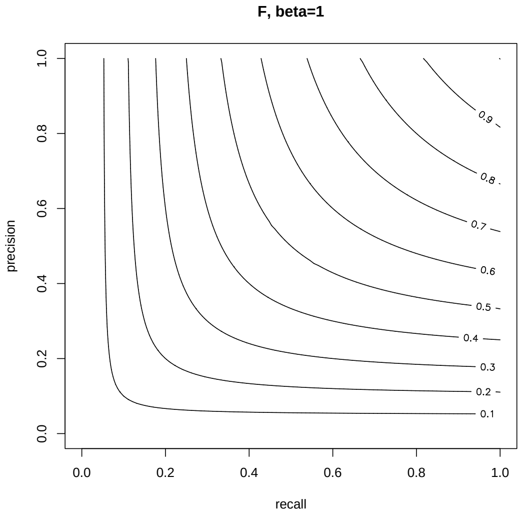


Figure 4: The F measure, plotted according to its precision and recall components. The lines show contours at intervals of 0.1 points of F.

ent 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 for novelty track runs is the F measure. This measure is a function of set recall and precision, together with a parameter β which determines the relative importance of recall and precision. A β value of 1, indicating equal weight, is used in the novelty track. $F_{\beta=1}$ is given as:

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

Alternatively, this can be formulated as

$$\mathrm{F} = \frac{2 \times (\text{\# relevant sentences retrieved})}{(\text{\# retrieved sentences}) + (\text{\# relevant sentences})}$$

For any choice of β, F lies in the range $[0, 1]$, and 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.

Note, however, that two runs with equal F scores do not indicate equal precision and recall. Figure 4 illustrates the shape of the F measure in precision-recall space. An F score of 0.5, for example, can reflect a range of precision and recall scores. Thus, two runs with equal F scores may be performing quite differently, and a difference in F scores can be due to changes in precision, recall, or both.

5 Participants

Table 2 lists the 14 groups that participated in the TREC 2003 novelty track. All but one group attempted the first task, and nearly every group tried every task. The rest of this section contains short summaries submitted by most of the groups about their approaches to the novelty task. For more details, please refer to the group's complete paper in the proceedings.

In general, most groups took a similar approach to the problem. Relevant sentences were selected by measuring similarity to the topic, and novel sentences by dissimilarty to past sentences. As can be seen from the following descriptions, there is a tremendous variation in how "the topic" and "past sentences" are modeled, and in how similarity is computed when sentences are involved. Many groups tried variations on term expansion to improve sentence similarity, some with more success than others.

5.1 CCS/University of Maryland [1]

For the 2003 DUC task of forming a summary based on the relevant and novel sentences, we tested a system based on a Hidden Markov Model (HMM). In this work, we use variations of this system on the tasks of the TREC Novelty Track. Our information retrieval system couples a query handler, a document clusterer, and a summary generator with a convenient user interface. Our summarization system uses an HMM to find relevant sentences in a document. The HMM has two types of states, corresponding to relevant and non-relevant sentences. The observation sequence scored by the HMM is composed of the number of signature terms and topic terms contained in each sentence. A signature term is a term that statistically occurs more frequently in the document set than in the document collection at large, and a subject term is a signature term which also occurs in the headlines or subject lines of a document. The counts of these terms are normalized within a document to have a mean of zero and variance of one. We determine the relevant sentences in a document based on

Table 2: Organizations participating in the TREC 2003 Novelty Track

		Runs submitted			
	Run prefix	Task 1	Task 2	Task 3	Task 4
Center for Computing Science / U. Maryland	ccsum	5	4	3	5
Chinese Academy of Sciences (CAS-ICT)	ICT	5	5	5	5
Chinese Academy of Sciences (CAS-NLPR)	NLPR	5	5	5	5
CL Research	clr	4	1	5	1
Indian Institute of Technology Bombay	IITB				1
Institut de Recherche en Informatique de Toulouse	IRIT	5	5		
LexiClone, Inc.	lexiclone	1			
Meiji University	Meiji	5	4	4	4
National Taiwan University	NTU	5	5	5	5
Tsinghua University	THU	5	3	4	5
University of Iowa	UIowa	2	5	2	5
University of Maryland Baltimore County	umbc	3	3		
University of Michigan	umich	5	5	5	5
University of Southern California-ISI	ISI	5			

the HMM posterior probability of each sentence being relevant. In particular, we choose the number of sentences to maximize the expected utility, which for TREC is simply the F1 score.

Several methods were explored to find a subset of the relevant sentences that has good coverage but low redundancy. In our multi-document summarization system, we used the QR algorithm on term-sentence matrices. For this work, we explored the use of the singular value decomposition as well as two variants of the QR algorithm.

5.2 Chinese Academy of Sciences (ICT) [11]

The novelty track can be treated as a binary classification problem: relevant sentences vs. irrelevant sentences, or new vs. non-new. In this way, we applied variants of techniques that have been employed for text categorization problem. To retrieve the relevant sentences, we compute the similarity between the topic and sentences using vector space model. The features for each topic are obtained by employing χ^2 statistic and each feature is also weighted using the χ^2 statistic. If the similarity exceeds a certain threshold, the sentence is considered as relevant. In addition, we try several techniques in an attempt to improve the performance. One is that the narrative section in the topic is analyzed to obtain the negative features and negative vector of the topic. We determine the relevance by adding similarity between the negative vector and sentence as a negative factor. The second, the threshold for different docs in each topic is dynamically adjusted according to the doc density, rather than fixed in the whole period. We have implemented the KNN algorithm and Winnow algorithm for classifying the sentences into relevant and irrelevant sentences in the novelty task 3. To detect the new sentences from the relevant sentences, we try several methods, such as Maximum Marginal Relevance (MMR) measure, Winnow algorithm and word overlapping within sentences. What's more, we attempt to detect novelty by computing semantic distance between sentences using WordNet.

5.3 Chinese Academy of Sciences (NLPR) [6]

For finding relevant sentences, we use a new statistical model called "Term Similarity Tree" to make the process of query expansion more flexible and controllable. Then, relevant feedback is used for additional modification for queries. Serveral different methods for similarity computing are developed to improve the performance. They are "simple window", "dynamic window", "active window". The key notion is that the window-based method can ensure that the closer the query words in sentences, the higher the similarity value. Finally, dynamic thresholds are used for different topics, which usually brings 1% increase of average F measure. For finding new sentences, We define a value called "New Information Degree" (NID) to present whether a sentence includes new information related to the former sentences. If the value of

NID is big, this sentence is reserved, or it will be discard. There are two different ways to define NID of the latter sentence related to the former sentence. One is based on idf value of terms and the other is based on bi-gram sequences.

5.4 CL Research [8]

The CL Research system parses and processes text into an XML representation, tagging the text with discourse, noun, verb, and preposition characteristics. The topic characterizations (titles, descriptions, and/or narratives) and the relevant documents provided by NIST were processed in this way. Componential analysis of the degree to which topic characterizations corresponded to sentences was used as the basis for determining relevance, using various scoring metrics. Similar componential analysis was used to compare each relevant sentence with all those that preceded it in order to assess novelty. Several variables were used as the basis for different runs under the different tasks (which also provided prior information that could be exploited), providing useful experimental results that will inform selection among alternatives for approaching the novelty task.

5.5 IRIT-SIG [2]

In TREC 2003, IRIT improved the strategy that was introduced in TREC 2002. A sentence is considered as relevant if it matches the topic with a certain level of coverage. This coverage depends on the category of the terms used in the texts. Three types of terms were defined for TREC 2002 highly relevant, lowly relevant and non-relevant (like stop words). In TREC 2003 we introduced a new class of terms: highly non-relevant terms. Terms from this category are extracted from the narrative parts of the queries that describe what will be a non-relevant document. A negative weight can be assigned to these words.

With regard to the novelty part, a sentence is considered as novel if its similarity with each of the previously processed and selected-as-novel sentences does not exceed a certain threshold. In addition, this sentence should not be too similar to a virtual sentence made of the n best-matching sentences.

5.6 University of Southern California-ISI

To identify opinion sentences, we used unigrams to indicate subjectivity. In addition to three baseline algorithms, we employed two sets of subjectivity-indicating words (either positive or negative valence, with appropriate strengths). One set was collected manually and extended with WordNet synonyms. The other was learned automatically from the Wall Street Journal. The words' relative scores and the algorithm's cutoff parameters were determined in a series of experiments. To our surprise the TREC results showed that one of our baselines (indicating that every sentence carries an opinion) actually beat the algorithm using the manually collected words. To identify event sentences, we adopted a standard IR procedure, treating each sentence as a separate document. For each event topic, we used all its non-stop words as query to extract event sentences. Again, the cutoff parameter was determined by experiment. We were happy to see that this method worked relatively well.

5.7 LexiClone [4]

For the sake of convenience we decided that on the word-per-word level, any language is about 58-59 percent nouns, 20 percent verbs and 20 percent adjectives. Except for prepositions, conjunctions, interjections, pronouns and other parts of speech that make up the remaining 1-2 percent, the rest of the language is a combination of these three dominant elements (or can be reduced to them). LexiClone establishes all possible combinations of nouns, verbs and adjectives for each sentence. We call these combinations "triads". (Actually, a triad is a smallest possible "key" phrase from a sentence.) After that we find sentences that have triads.

5.8 Meiji University [9]

For identifying relevant sentences, we employed following information-filtering-based approach. We regarded sentences as very short documents. Initial profiles, which are made from topic descriptions, are expanded conceptually. Conceptual fuzzy sets, which we proposed previously, are used for conceptual expansion. If the cosine similarity between the expanded profile and a word vector of each sentence exceeds a threshold, the sentence is regarded as relevant. For identifying new sentences, we considered two measures; sentence score and redundancy score. 1) For calculating a sentence score, we used N-window-idf as a time window. Local sentence score is calculable by using document frequency of past N documents. 2) Redundancy score is the maximum value of the similarity with the sentence judged to be novel in the past.

5.9 National Taiwan University [12]

According to the results of TREC 2002, we realized the major challenge issue of recognizing relevant sentences is a lack of information used in similarity computation among sentences. In TREC 2003, NTU attempts to find relevant and novel information based on variants of employing information retrieval (IR) system. We call this methodology IR with reference corpus, which can also be considered an information expansion of sentences. A sentence is considered as a query of a reference corpus, and similarity between sentences is measured in terms of the weighting vectors of document lists ranked by IR systems. Basically, we looked for relevant sentences by comparing their results on a certain information retrieval system. Two sentences are regarded as similar if they are related to the similar document lists returned by IR system. In novelty parts, similar analysis is used to compare each relevant sentence with all those that preceded it to find out novelty. An effectively dynamic threshold setting which is based on what percentage of relevant sentences is within a relevant document is presented.

5.10 Tsinghua University [14]

Research in IR group of Tsinghua University on this year's novelty track mainly focused on four aspects: (1) unsupervised relevance judgment, where QE and pseudo relevance feedback has been used. (2) efficient sentence redundancy computing: we used unsymmetrical sentence "overlap" metric, sub-topic redundancy elimination and sentence clustering. (3) supervised sentence classification, where a SVM classifier has been used and got encouraging results; (4) supervised redundancy threshold learning. A new IR system named TMiner has been built on which all experiments have been performed.

5.11 University of Iowa [3]

Our approach is basically the same as that used last year. We use new named entity and noun phrase triggering, guarded by a dual threshold of sentence similarity and full-document similarity. If the full document is sufficiently similar and the current sentence is sufficiently similar, the number of newly-detected named entities and noun phrases is compared against a minimum threshold and if the minimum is met, the current sentence is declared to be novel. The named entities used include persons, organizations and place names. Relevance is simple term similarity.

5.12 University of Maryland Baltimore County [7]

To find the relevant sentences, we used a method comprising of query expansion and sentence clustering. In the query expansion step, we experimented with two methods, one was to determine highly cooccurring terms by means of a SVD analysis and, the other was by determining meaningful terms as obtained by a language analysis of the narrative section for each topic. The sentences, per topic, were clustered and the top clusters were selected based on similarity scores of the cluster centroids and the expanded query. All the sentences from the selected clusters are chosen as the relevant sentences.

To find the novel sentences, we experimented with two methods. One, based on a text summarization method, was clustering relevant sentences and choosing one sentence each from the selected clusters to make up the set of novel sentences. In the second method, using a sentence-sentence similarity matrix (of relevant sentences), the dissimilarity between sentences was used to determine novel sentences.

5.13 University of Michigan [10]

First we used the MEAD summarization software to compute scores for each sentence on features such as length, position, word overlap with query, title and description. Since we trained maximum entropy classifiers, these scores were then discretized. Once the MEAD features were calculated, discretized and formatted, we used the maxent-2.1.0 software to train our models for novel and relevant sentences.

For tasks 1 and 3, once the maxent models had been trained for classifying novel and relevant sentences and were used to produce a ranked list of sentences as to how likely they were to be novel or relevant, we then chose differing percentage cut offs for each run in an attempt to maximize recall and precision on our devtest data set. For tasks 2 and 4, we noted that the F-measure for a baseline algorithm of submitting all relevant sentences as being novel was quite high. Therefore, we focused on trying various discretizations of our feature scores in order to improve the classifier's performance on the devtest set

6 Results

Figures 5, 7, 8, and 9 show the average F scores in each task. Task 1 scores are shown alongside the "scores" of the secondary assessor, who may be considered to have been performing this task. Within the margin of error of human disagreement, these lines can be thought of as representing the best possible performance. The best systems are performing at this level. Nine runs have novelty F scores of 0 because those runs did not return any novel sentences.

Tasks 1 and 3 show novelty retrieval performance closely tracking relevant retrieval performance. Only a few runs near the bottom of the performance range did better at retrieving novel sentences than relevant ones. This seems somewhat surprising, since while the retrieved set of relevant sentences places a bound on recall for the novel set (since only retrieved sentences can be labeled novel), any level of precision is possible, and thus there isn't any reason why F_{novel} shouldn't exceed $F_{relevant}$. However, to achieve this most systems would have had to make a very large improvement in precision when retrieving novel sentences.

As stated previously, sometimes it can be hard to understand what the F score means in terms of the actual behavior of each run. Figure 6 shows the F scores for task 1, along with each run's corresponding average recall and precision. Note for example the run ISIALL03 (run #11 on the x axis), which retrieved only relevant sentences, and retrieved all of them; for this run, average recall was 1.0 but precision was 0.41. It is very interesting to note that average recall seems to correlate more closely to the F scores, although F is defined to be a harmonic mean between the two. This may mean that within each run, recall was more consistent across topics than was precision.

The scores for tasks 2–4 show how many of the systems can take advantage of training data, both for relevance and novelty. Comparing the graph of tasks 2 and 3, we can see that having more relevance information dramatically improves novelty retrieval effectiveness. Moreover, comparing tasks 2 and 4, we can see that having relevant sentences is more valuable than having novel sentences for training, since the top systems do not improve from task 2 to task 4.

The graphs for tasks 2 and 4 compare the runs against a baseline system which merely returns all the relevant sentences (provided as training data in these tasks) as novel. The best systems are performing above this baseline, indicating that they are being somewhat selective in what they return as novel.

Event topics were easier than opinion topics. Figure 10 illustrates this phenomenon in task 1. Relevant sentence retrieval scores are on the left, novelty retrieval scores on the right. The graphs show the overall average along with the averages for event and opinion topics for each run. Nearly every run did better at events than opinions; the exceptions are UMBC and NTU for relevant sentences, and NTU and one IRIT run for novel sentences.

As the systems receive more relevant sentences as training data, they improve on opinion topics. In task 3 (where systems received some relevant and novel training data), all systems perform as well or better on event topics than on opinions. However, in tasks 2 and 4, where the systems receive complete relevance information, the situation is reversed: all systems do better on opinion topics. Clearly, the systems are less able to identify relevant sentences in opinion topics, but if they know which ones are relevant, they do better on opinion topics than on events. Having a small amount of relevant sentence training data (as in task 3) is not sufficient to boost a system's overall performance.

References

[1] J. M. Conroy, D. M. Dunlavy, and D. P. O'Leary. From TREC to DUC to TREC again. In Voorhees and Harman [13].

[2] T. Dkaki and J. Mothe. TREC novelty track at IRIT-SIG. In Voorhees and Harman [13].

[3] D. Eichmann, P. Srinivasan, M. Light, H. Wang, X. Y. Qiu, R. J. Arens, and A. Sehgal. Experiments in novelty, genes and questions at the University of Iowa. In Voorhees and Harman [13].

[4] I. S. Geller. The role and meaning of predicative and non-predicative definitions in the search for information. In Voorhees and Harman [13].

[5] Donna Harman. Overview of the TREC 2002 novelty track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, NIST Special Publication 500-251, pages 46–55, Gaithersburg, MD, November 2002.

[6] Q. Jin, J. Zhao, and B. Xu. NLPR at TREC 2003: Novelty and robust. In Voorhees and Harman [13].

[7] S. Kallukar, Y. Shi, R.S. Cost, C. Nicholas, A. Java, C. James, S. Rajavaram, V. Shanbhag, S. Bhatkar, and D. Ogle. UMBC at trec 12. In Voorhees and Harman [13].

[8] K. C. Litkowski. Use of metadata for question answering and novelty tasks. In Voorhees and Harman [13].

[9] R. Ohgaya, A. Shimmura, and T. Takagi. Meiji university web and novelty track experiements at TREC 2003. In Voorhees and Harman [13].

[10] J. Otterbacher, H. Qi, A. Hakim, and D. Radev. The University of Michigan at TREC 2003. In Voorhees and Harman [13].

[11] J. Sun, Z. Yang, W. Pan, H. Zhang, B. Wang, and X. Cheng. TREC 2003 novelty and web track at ICT. In Voorhees and Harman [13].

[12] M.-F. Tsai, M.-H. Hsu, and H.-H. Chen. Approach of information retrieval with reference corpus to novelty detection. In Voorhees and Harman [13].

[13] E. M. Voorhees and D. K. Harman, editors. *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, MD, 2003.

[14] M. Zhang, C. Lin, Y. Liu, L. Zhao, and S. Ma. THUIR at TREC 2003: Novelty, robust and web. In Voorhees and Harman [13].

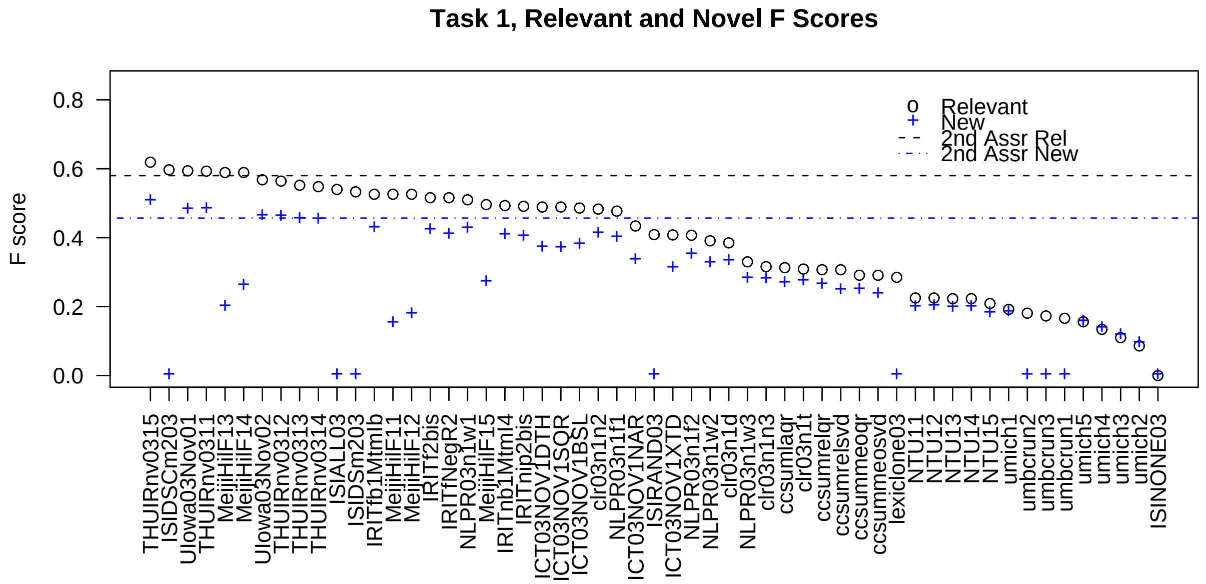


Figure 5: Scores for Task 1, along with the "average score" of the secondary assessor.

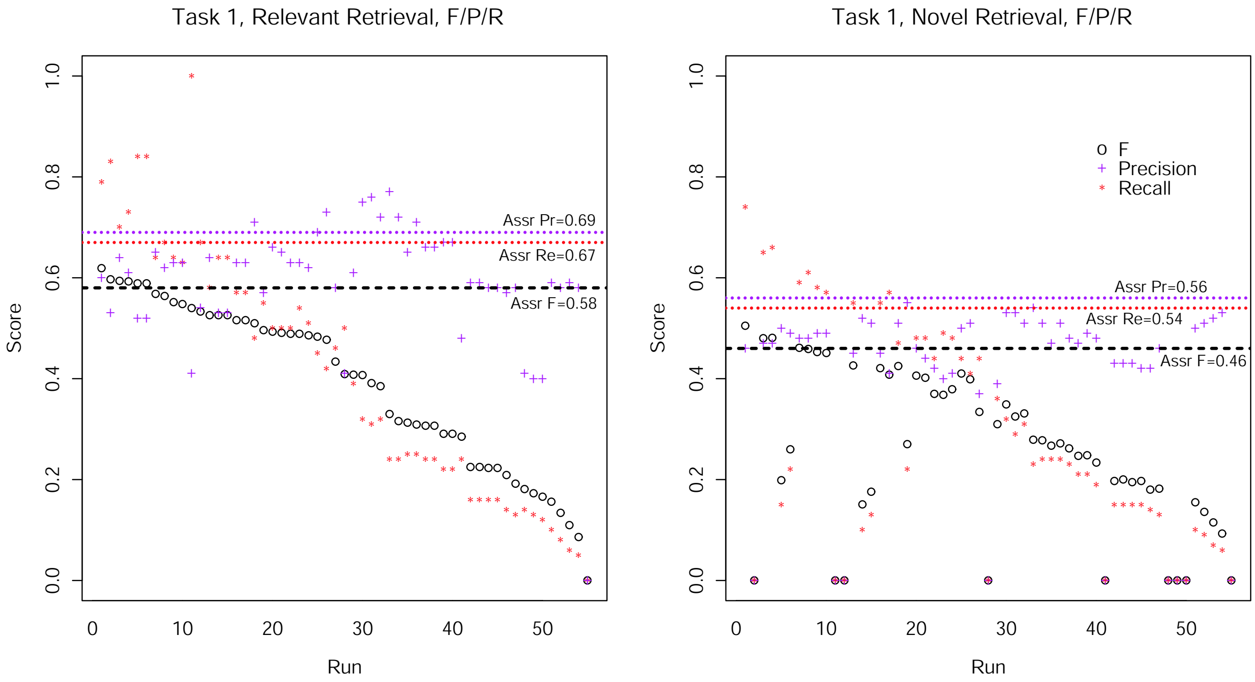


Figure 6: Task 1 relevant and novel F scores, with corresponding precision and recall.

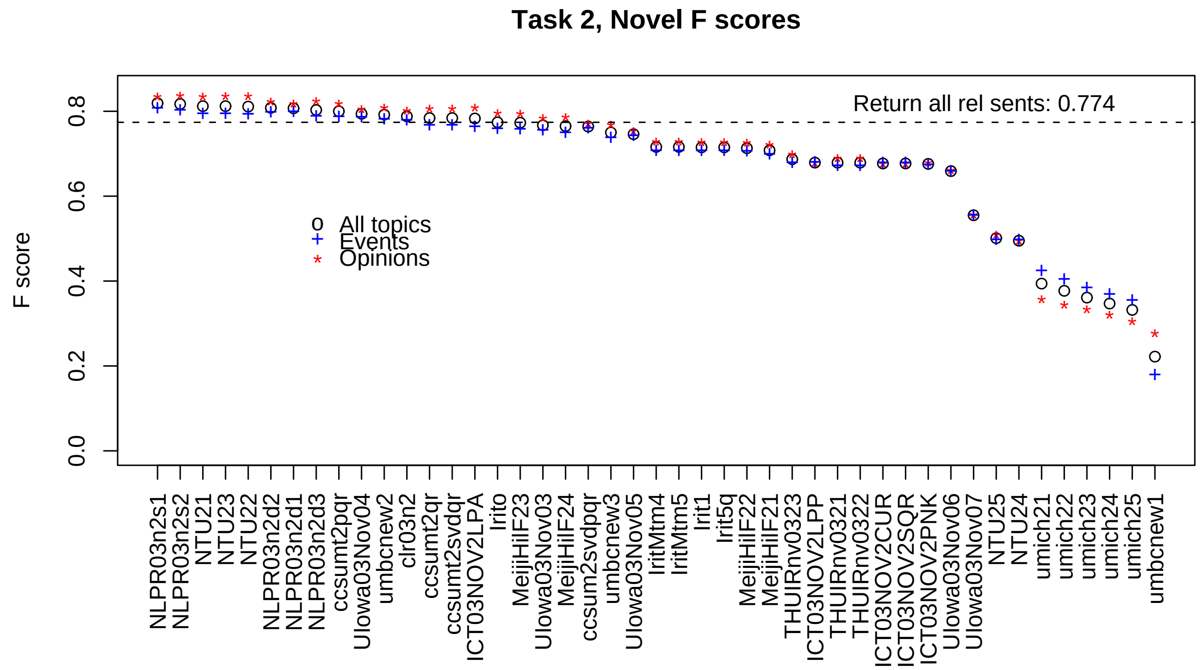


Figure 7: Scores for Task 2, against a baseline of returning all relevant sentences as novel.

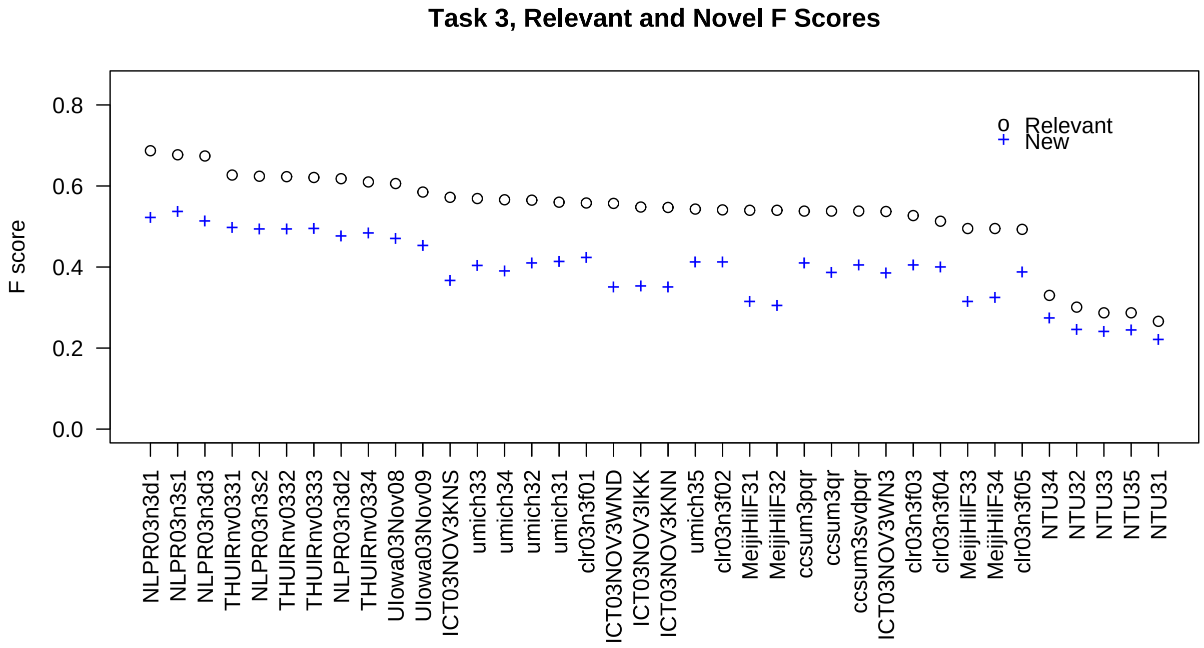


Figure 8: Scores for Task 3.

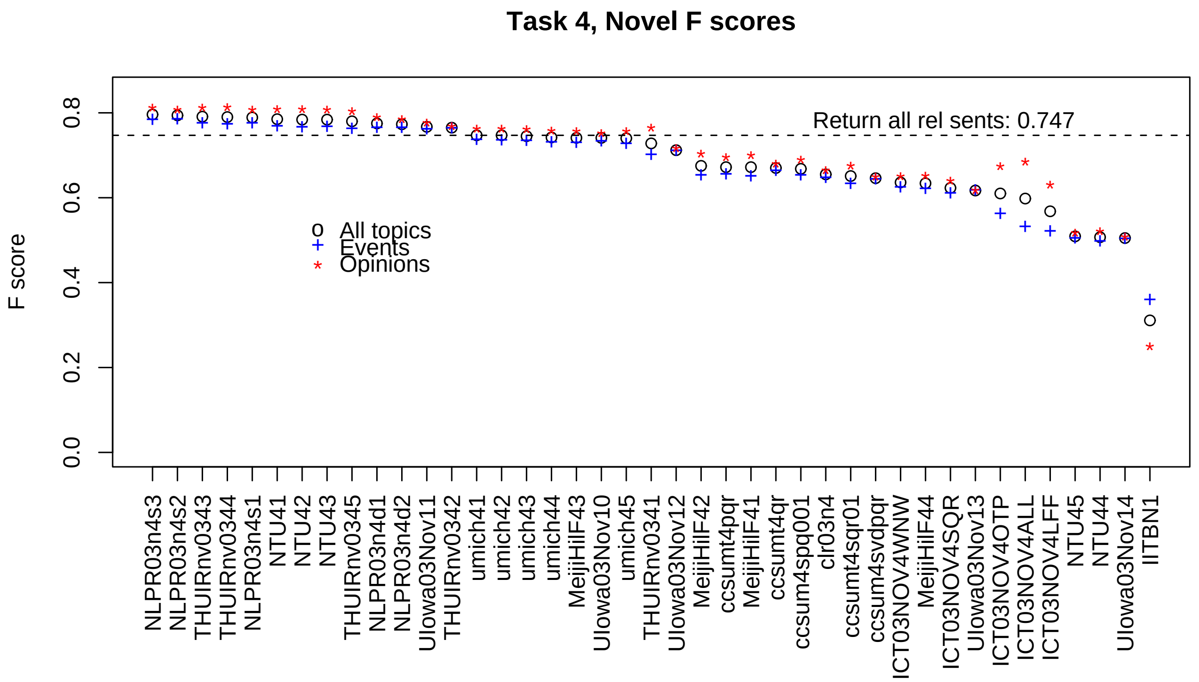


Figure 9: Scores for Task 4, against a baseline of returning all relevant sentences as novel.

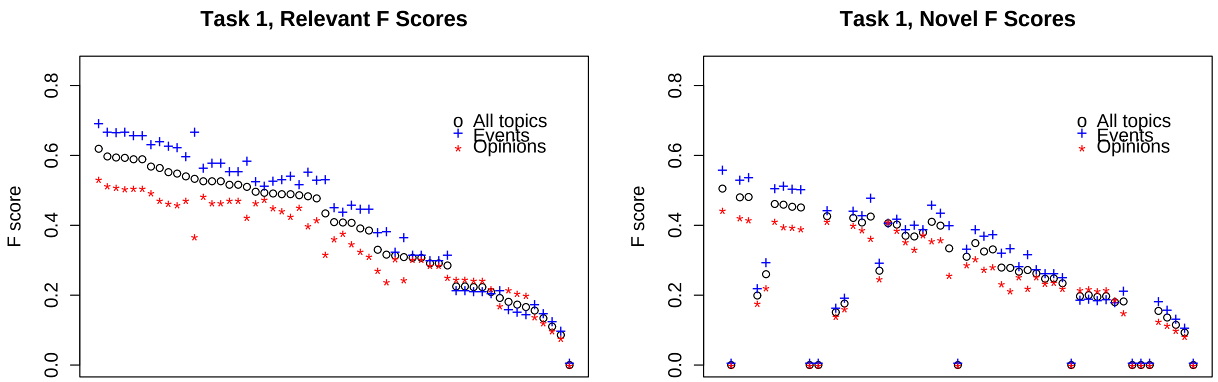


Figure 10: Scores for task 1, broken down by topic type. Runs are along the X axis; the run names have been omitted for readability, but the runs are in the same order as in Figure 5.