Overview of the TREC 2004 Novelty Track

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

Abstract

TREC 2004 marks the third and final year for the novelty track. The task is as follows: 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. As in 2003, there were two categories of topics – events and opinions – and four subtasks which provided systems with varying amounts of relevance or novelty information as training data. This year, the task was made harder by the inclusion of some number of irrelevant documents in document sets. Fourteen groups participated in the track this year.

1 Introduction

The novelty track was introduced in TREC 2002 [1]. The basic task is as follows: given a topic and an ordered set of 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, the novelty track has four different 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.

The track has changed slightly from year to year. The first run in 2002 used old topics and relevance judgments, with sentences judged by new assessors [1]. TREC 2003 included separate tasks, made the document ordering chronological rather than relevance-based, and introduced new topics and the different topic types [2]. This year, the major change is the inclusion (or perhaps re-introduction) of irrelevant documents into the document sets.

2 Input Data

The documents for the novelty track are 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). As a result, this collection exhibits greater redundancy than other TREC collections, and thus less novel information, increasing the realism of the task.

The NIST assessors created fifty new topics for the 2004 track. As was done last year, the topics were of two types. Twenty-five topics concerned events, such as India and Pakistan's nuclear tests in 1998, and twenty-five topics focused on opinions about controversial subjects such as the safety of irradiated food and the so-called "abortion pill" RU-486. The topic type was indicated in the topic description by a `<toptype>` tag. The assessors, in creating their topics, searched the AQUAINT collection using WebPRISE, NIST's IR system, and collected 25 documents they deemed to be relevant to the topic. They also labeled some documents as irrelevant, and all documents judged irrelevant and ranked above the 25 relevant documents were included in the document sets. Note that this means that the irrelevant documents are close matches to the relevant ones, and not random irrelevant documents.

Once selected, the documents were ordered chronologically. (Chronological ordering is achieved trivially in the AQUAINT collection by sorting document IDs.) 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

There are four tasks in the novelty track:

Task 1. Given the set of documents for the topic, identify all relevant and novel sentences.

Task 2. Given the relevant sentences in all 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 documents. Note that since some documents are irrelevant, there *may not be* any relevant or novel sentences in the first 5 documents for some topics.

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

The assessors only judged sentences in the relevant documents. Since, by the definition of relevance in TREC, a document containing any relevant information would itself be relevant, the assessors would not miss any relevant information by not judging the sentences in the irrelevant documents. This does give the second assessor some advantage against systems attempting task 1, since the assessor was not confronted with irrelevant documents in the sentence judging phase.

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. The first novelty track topics (in 2002) were created using topics from old TRECs and relevant documents from manual TREC runs, and the sentences judgments were made by new assessors. Those topics had very few relevant sentences and consequently nearly every relevant sentence was novel. Last year's topics, which were each newly developed and judged by a single assessor, resulted in topics with much more reasonable levels of relevant and new information. This year the inclusion of irrelevant documents means that fewer sentences are relevant. Somewhat surprisingly, perhaps, the fraction of relevant sentences which are novel is lower than last year as well.

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

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
N51	E	669	C	107	15.99	26	24.30	B	112	16.74	38	33.93
N53	E	667	E	106	15.89	31	29.25	C	136	20.39	86	63.24
N54	E	1229	E	198	16.11	71	35.86	B	384	31.24	224	58.33
N55	E	536	C	56	10.45	21	37.50	E	96	17.91	46	47.92
N56	E	1904	E	196	10.29	103	52.55	A	133	6.99	47	35.34
N57	E	378	B	21	5.56	10	47.62	D	170	44.97	116	68.24
N59	E	855	D	214	25.03	86	40.19	C	152	17.78	62	40.79
N64	E	679	C	214	31.52	140	65.42	A	228	33.58	64	28.07
N68	E	1331	B	200	15.03	45	22.50	E	210	15.78	82	39.05
N69	E	367	D	169	46.05	55	32.54	B	122	33.24	59	48.36
N72	E	1007	B	147	14.60	43	29.25	D	144	14.30	48	33.33
N73	E	380	D	268	70.53	139	51.87	A	164	43.16	93	56.71
N74	E	502	D	240	47.81	107	44.58	C	129	25.70	79	61.24
N79	E	1580	C	199	12.59	69	34.67	D	188	11.90	116	61.70
N80	E	447	E	74	16.55	48	64.86	B	104	23.27	51	49.04
N81	E	684	A	173	25.29	31	17.92	C	236	34.50	167	70.76
N82	E	1152	C	355	30.82	165	46.48	B	100	8.68	44	44.00
N83	E	816	A	250	30.64	62	24.80	E	227	27.82	122	53.74
N85	E	1419	B	181	12.76	95	52.49	E	116	8.17	59	50.86
N87	E	1026	D	476	46.39	163	34.24	C	369	35.96	231	62.60
N88	E	708	C	312	44.07	171	54.81	E	307	43.36	131	42.67
N90	E	1971	B	529	26.84	168	31.76	D	762	38.66	310	40.68
N92	E	879	B	188	21.39	172	91.49	A	199	22.64	83	41.71
N95	E	627	E	78	12.44	36	46.15	D	168	26.79	108	64.29
N98	E	408	C	171	41.91	65	38.01	A	267	65.44	67	25.09
N52	O	1018	B	103	10.12	55	53.40	C	298	29.27	202	67.79
N58	O	1346	A	146	10.85	42	28.77	C	252	18.72	163	64.68
N60	O	948	B	172	18.14	64	37.21	A	257	27.11	79	30.74
N61	O	1150	A	70	6.09	21	30.00	B	78	6.78	40	51.28
N62	O	3132	E	89	2.84	45	50.56	D	97	3.10	79	81.44
N63	O	518	B	49	9.46	21	42.86	E	84	16.22	55	65.48
N65	O	705	B	95	13.48	61	64.21	C	113	16.03	90	79.65
N66	O	795	A	195	24.53	25	12.82	E	286	35.97	137	47.90
N67	O	423	E	113	26.71	72	63.72	C	109	25.77	82	75.23
N70	O	1030	D	94	9.13	31	32.98	E	237	23.01	104	43.88
N71	O	908	B	62	6.83	28	45.16	A	127	13.99	28	22.05
N75	O	2922	B	169	5.78	100	59.17	C	284	9.72	245	86.27
N76	O	1697	A	217	12.79	51	23.50	D	118	6.95	39	33.05
N77	O	1144	D	74	6.47	23	31.08	B	102	8.92	36	35.29
N78	O	1308	A	145	11.09	59	40.69	B	59	4.51	25	42.37
N84	O	1363	D	101	7.41	31	30.69	E	153	11.23	80	52.29
N86	O	493	D	67	13.59	33	49.25	A	96	19.47	46	47.92
N89	O	1271	B	204	16.05	130	63.73	A	181	14.24	61	33.70
N91	O	1473	B	112	7.60	51	45.54	D	123	8.35	99	80.49
N93	O	1017	B	181	17.80	56	30.94	E	255	25.07	129	50.59
N94	O	1099	E	102	9.28	59	57.84	A	91	8.28	46	50.55
N96	O	1328	A	131	9.86	60	45.80	D	61	4.59	45	73.77
N97	O	1416	A	123	8.69	31	25.20	B	122	8.62	89	72.95
N99	O	1192	C	259	21.73	131	50.58	D	495	41.53	341	68.89
N100	O	530	E	148	27.92	52	35.14	B	152	28.68	78	51.32

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**).

Because this year’s document sets include irrelevant documents, the fraction of relevant sentences is less than half that of last year: a mean of 19.2%, compared with 41.1% in TREC 2003. However, the amount of novel information as a fraction of relevant is also lower: a 42% this year vs. 64.6% in TREC 2003. This was somewhat surprising as the collection and topic types are the same, and the topics have the same number of relevant documents. Beyond simple intertopic variation, these topics just have more redundant information.

Opinion topics tended to have fewer relevant sentences than event topics. 25.9% of sentences in event topics were relevant, compared to only 12.6% in opinion topics. Even though the topics are about opinions, the documents are still news stories and thus include current events and background information in addition to the relevant opinion material. The fraction of relevant sentences which were novel was the same for both types, 42%.

In examining assessor effects, this year we were able to achieve much better balance in the second round of assessing, with each assessor judging five topics written by someone else. Overall, the assessors tended to find about the same amount of relevant information whether they were judging their own topics or someone else’s (19.2% for their own topics vs. 21.7% in the second round, not significant by a t-test), but identified more novel sentences (42% vs. 52.6%, significant at $p = 0.0009$). We have not made a detailed analysis of how the assessors differed in particular judgments or in their judging patterns.

In summary, the topics for this year seem comparable in quality to the TREC 2003 topics, with minimal assessor effects. The inclusion of irrelevant documents makes the task this year harder for systems, and thus the two topic sets should not be combined.

4.2 Scoring

The sentences selected manually by the NIST assessor who created the topic were considered the truth data. The judgments by the secondary assessor were taken as a human baseline performance in the first task. Relevant and novel sentence retrieval have each been evaluated separately.

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. One alternative is to use set-based recall and precision. 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 the TREC 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 set recall and set precision averaged over a set of topics is not a robust diagnostic indicator of system performance. There is also the problem of how to define precision when the system returns no sentences ($S = 0$). Leaving that topic out of the evaluation for that run would mean that different systems would be evaluated over different numbers of topics, while defining precision in the degenerate case 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. The F measure (from van Rijsbergen’s E 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

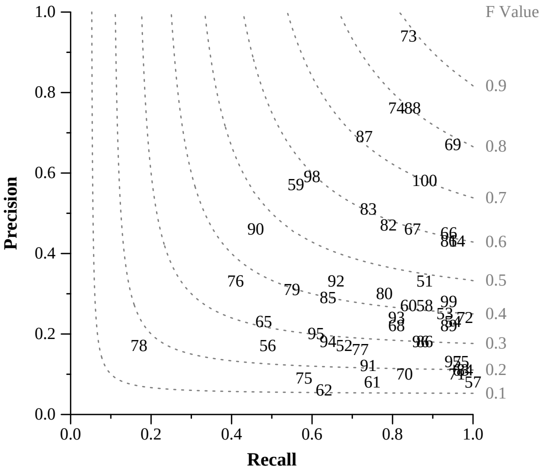

Figure 1: The F measure, plotted according to its precision and recall components. The lines show contours at intervals of 0.1 points of F. The black numbers are per-topic scores for one novelty track run.

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

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 1 illustrates the shape of the F measure in precision-recall space. An F score of 0.5, for example, can describe a range of precision and recall scores. Figure 1 also includes the per-topic scores for a particular run are also plotted. It is easy to see that topics 98, 83, 82, and 67 exhibit a wide range of performance, but all have an F score of close to 0.6. 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 2004 novelty track. Nearly every group attempted the first two tasks, but tasks three and four were less popular than last year, with only 8 groups participating in each (compared to 10 last year). 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.

Most groups took a similar high-level approach to the problem, and the range of approaches is not dramatically different from last year. Relevant sentences were selected by measuring similarity to the topic, and novel sentences by dissimilarity 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, how similarity is computed when sentences are involved, and what constitutes the thresholds for relevance and novelty. Many groups tried variations on term expansion to improve sentence similarity, some with more success than others.

5.1 Chinese Academy of Sciences – ICT

In TREC 2004, ICT divided novelty track into four sequential stages. It includes: customized language parsing on original dataset, document retrieval, sentence relevance and novelty detection. In the first preprocessing stage, we applied sentence segmenter, tokenization, part-of-speech tagging, morphological analysis, stop word remover and query analyzer on topics and documents. As for query analysis, we categorized words in topics into description words and content words. Title, description and narrative parts are all merged into query with different weights. In the stage of document and sentence retrieval, we introduced vector space model (VSM) and its variants, probability model (OKAPI) and statistical language model. Based on VSM, we tried various query expansion strategies: pseudo-feedback, term expansion with synset or synonym in WordNet and expansion with highly local co-occurrence terms. With regard to the novelty stage, we defined three types of new degree: word overlapping and its extension, similarity comparison and information gain. In the last three tasks, we used the known results to adjust threshold, estimate the number of results, and turn to classifier, such as inductive and transductive SVM.

5.2 CL Research

The CL Research novelty assessment is based on a full-scale parsing and processing of documents and

Table 2: Organizations participating in the TREC 2004 novelty track

		Runs submitted			
	Run prefix	Task 1	Task 2	Task 3	Task 4
Chinese Academy of Sciences (CAS-ICT)	ICT	5	5	4	5
CL Research	clr	2	1	4	1
Columbia University	nov		5		
Dublin City University	cdvp	5	5		
IDA / Center for Computing Science	ccs	5	5	4	
Institut de Recherche en Informatique de Toulouse	IRIT	5	2	5	
Meiji University	Meiji	3	5	3	5
National Taiwan University	NTU	5	5		
Tsinghua University	THUIR	5	5	5	5
University of Iowa	UIowa	5	5	5	5
University of Massachusetts	CIIR	2	5	3	
University of Michigan	umich	5	5	5	4
Université Paris-Sud / LRI	LRI	5	5		
University of Southern California-ISI	ISI	5			

topic descriptions (titles, descriptions, and narratives) into an XML representation characterizing discourse structure, syntactic structure (particularly noun, verb, and prepositional phrases), and semantic characterizations of open-class words. Componential analysis of the topic narratives was used as the basis for identifying key words and phrases in the document sentences. Several scoring metrics were used to determine the relevance for each sentence. In TREC 2004, the presence of communication nouns and verbs in the narratives was used to expand relevance assessments, by identifying communication verbs in the sentences. This significantly increased recall over TREC 2004, without a significant degradation of precision. CL Research's novelty component was unchanged, but precision on Task 2 was considerably lower. This lower precision was observed in other tasks as well, and perhaps reflects the significantly lower scores among all participants. CL Research has set up an evaluation framework to examine the reasons for these lower scores.

5.3 Columbia University

Our system for the novelty track at TREC 2004, SumSeg, for Summary Segmentation, is based on our observations of data we collected for the development of our system to prepare update summaries, or bulletins. We see that new information often appears in text spans of two or more sentences, and at other times, a piece of new information is embedded within a sentence mostly containing previously seen material. In order to capture both types of cases, we avoided direct sentence similarity measures, and took evidence of unseen words as evidence of novelty. We employed a hill climbing algorithm to learn thresholds for how many new words would trigger a novel classification. We also sought to learn different weights for different types of nouns, for example, persons, or locations or common nouns. In addition, we included a mechanism to allow sentences that had few strong content words to ”continue” the classification of the previous sentence. Finally, we used two statistics, derived from analysis of the full AQUAINT corpus, to eliminate low-content words. We submitted a total of five runs: two used learned parameters to aim at high precision output, and one aimed at higher recall. Another run was a straightforward vector-space model used as a baseline, and the last was a combination of the high recall run with the vector-space model. Training was done on the 2003 TREC novelty data.

5.4 Dublin City University

This is the first year that DCU has participated in the novelty track. We built three models; the first focused on retrieving the twenty-five documents that were relevant to each topic; the second focused on retrieving relevant sentences from this list of retrieved documents to satisfy each individual topic; the third focused on the detection of novel sentences from this relevant list. In Task1 we used an information retrieval system developed by the CDVP for the terabyte track as a basis for our experiments. This system used the BM25 ranking algorithm. We used various query and document expansion techniques to enhance the performance for sentence level retrieval. In Task 2 we developed two formulas, the ImportanceValue and The NewSentenceValue, which exploit term characteristics using traditional document similarity methods.

5.5 Institut de Recherche en Informatique de Toulouse (IRIT)

In TREC 2004, IRIT modified important features of the strategy that was developed for TREC 2003. These features include both some parameter values, topic expansion and taking into account the order of sentences. According to our method, a sentence is considered 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. Four types of terms have been defined — highly relevant, scarcely relevant, non-relevant (like stop words), highly non-relevant terms (negative terms). Term categorization is based on topic analysis: highly non-relevant terms are extracted from the narrative parts that describe what will be a non-relevant document. The three other types of terms are extracted from the rest of the query and are distinguished according to the score they obtain. The score is based both on the term occurrence and on the topic part they belong to (Title, descriptive, narrative). Additionally we increase the score of a sentence when the previous sentence is relevant. When topic expansion is applied, terms from relevant sentences (task 3) or from the first retrieved sentences (task 1) are added to the initial terms. 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 Iowa

Our system for novelty this year comprises three distinct variations. The first is a refinement of that used for last year involving named entity occurrences and functions as a comparative baseline. The second variation extends the baseline system in an exploration of the connection between word sense and novelty through two alternatives. The first alternative attempts to address the semantics of novelty by expanding all noun phrases (and contained nouns) to their corresponding WordNet synset IDs, and subsequently using synset IDs for novelty comparisons. The second alternative performs word sense disambiguation using an ensemble scheme to establish whether the additional computational overhead is warranted by an increase in performance over simple sense expansion.

The third variation involves more 'traditional' similarity schemes in the positive sense for relevance and the negative sense for novelty. SMART is first used to identify the top 25 documents and then judges relevance at the sentence level to generate a preliminary pool of candidates and then incrementally extends a matched terminology vector. The matched term vector is then used to rematch candidate sentences. Only similarities below a threshold - and hence possessing sufficient dissimilarity are declared novel.

5.7 University of Massachusetts

For relevant sentences retrieval, our system treated sentences as documents and took the words in the title field of the topics as queries. TFIDF techniques with selective feedback were used for retrieving relevant sentences. Selective pseudo feedback means pseudo feedback was performed on some queries but not on other queries based on an automatic analysis on query words across different topics. Basically, a query with more focused query words that rarely appear in relevant documents related to other queries was likely to have a better performance without pseudo feedback. Selective relevance feedback was performed when relevance judgment of top five documents was available as for Task 3. Whether to performance relevance feedback on a query was determined by the comparison between the performance with and without relevance feedback in the top five documents for this query.

For identifying novel sentences, our system started with the sentences returned from the relevant sentences retrieval. The cosine similarity between a sentence and each previous sentence was calculated. The maximum similarity was used to eliminate redundant sentences. Sentences with a maximum similarity greater than a preset threshold were treated as redundant sentences. The value of the same threshold for all topics was tuned with the TREC 2003 track data when no judgment was available. The value of the threshold for each topic was trained with the training data when given the judgment of the top five documents. In addition to the maximum similarity, new words and named entities were also considered in identifying novel sentences.

5.8 University of Michigan

We view a cluster of documents as a graph, where each node is a sentence. We define an edge between two nodes where the cosine similarity between the corresponding two sentences is above a predefined threshold. After this graph is constructed, we find the eigenvector centrality score for each sentence by using a power method, which also corresponds to the stationary distribution of the stochastic graph.

To find the relevant sentences, we compute eigenvector centrality for each sentence together with some other heuristic features such as the similarity between the sentence and the title and/or description of the topic. To find the new sentences, we form the cosine similarity graph that consists of only the relevant sentences. Since the order of the sentences is important, unlike the case in finding the relevant sentences, we form a directed graph where every sentence can only point to the sentences that come after and are similar to it. The more incoming edges a sentence has, the more repeated information it contains. Therefore, the sentences with low centrality scores are considered as new. The system is trained on 2003 data using maximum entropy or decision lists.

5.9 Université Paris-Sud – LRI

The text-mining system we are building deals with the specific problem of identifying the instances of relevant concepts found in the texts. This has several consequences. We develop a chain of linguistic treatment such that the n-th module improves the semantic tagging of the $(n-1)$-th. This chain has to be friendly toward at least two kinds of experts: a linguistic expert, especially for the modules dealing mostly with linguistic problems (such as correcting wrong grammatical tagging), and a field expert for the modules dealing mostly with the meaning of group of words. Our definition of friendliness includes also developing learning procedures adapted to various steps of the linguistic treatment, mainly for grammatical tagging, terminology, and concept learning. In our view, concept learning requires a special learning procedure that we call Extensional Induction. Our interaction with the expert differs from classical supervised learning, in that the expert is not simply a resource who is only able to provide examples, and unable to provide the formalized knowledge underlying these examples. This is why we are developing specific programming languages which enable the field expert to intervene directly in some of the linguistic tasks. Our approach is thus not particularly well adapted to the TREC competition, but our results show that the whole system is functional and that it provides usable information.

In this TREC competition we worked at two levels of our complete chain. In one level, we stopped the linguistic treatment at the level of terminology (i.e., detecting the collocations relevant to the text). Relevance was then defined as the appearance of the same terms in the task definition (exactly as given by the TREC competition team) and in the texts. Our relatively poor results show that we should have been using relevance definitions extended by human-provided comments. Novelty was defined by a TF*IDF measurement which seems to work quite correctly, but that could be improved by using the expert-defined concepts as we shall now see. The second level stopped the linguistic treatment after the definition of the concepts. Relevance was then defined as the presence of a relevant concept and novelty as the presence of a new concept. For each of the 5 runs, this approach proved to be less efficient than the simpler first one. We noticed however that the use of concepts enabled us to obtain excellent results on specific topics (and extremely bad ones as well) in different runs. We explain these very irregular results by our own lack of ability to define properly the relevant concepts for all the 50 topics since we got our best results on topics that either we understood well (e.g., Pinochet, topic N51) or that were found interesting (e.g., Lt-Col Collins, topic N85).

5.10 University of Southern California – ISI

Our system's two modules recognize relevant event and opinion sentences respectively. We focused mainly on recognizing relevant opinion sentences using various opinion-bearing word lists. This year, each topic contained 25 relevant documents, possibly mixed with additional irrelevant documents. Thus, before proceeding to the next phase we had to separate relevant documents from irrelevant documents. We treat this problem as a standard Information Retrieval (IR) procedure. We used a probabilistic Bayesian inference network model to identify the relevant documents. For opinion topics, we used unigrams as subjectivity clues and built four different systems to generate opinion-bearing word lists. After building these unigram lists, we checked each sentence in the relevant documents for the presence of opinion-bearing words. For event topics, we treat event identification as a traditional document IR task. For the IR part we treat each sentence independently of other sentences and index them accordingly. We thus reduce the problem of event identification to that of sentence retrieval. We choose the description `<desc>` field for formulating the query.

5.11 Tsinghua University

- Text feature selection and reduction, including using Named Entities, POS-tagging information, and PCA transformation which has been shown to be more effective;
- Improve sentence classification to find relevant information using SVM;
- Efficient sentence redundancy computing, including selected pool approach, tightness restriction factor, and PCA-based cosine similarity measurement;
- Effective result filtering, combining sentence and document similarities.

Several approaches are investigated for the step two of novelty (redundancy reduction): Combining the pool method and sentence to sentence overlap, we have a selected pool method, where unlike in the pool method, not all previously seen sentences are included into the pool, only those thought to be related are included. Tightness restriction to overcome one disadvantage of overlap methods is studied. We observed not all sentences with an overlap of 1 (complete term overlap) are really redundant, so we came up with the idea of tightness restriction which tries to recover highly overlapping but in fact novel sentences. In this method, the ratio of the range of common terms in the previous sentence over the range in the later sentence is used as a statistic. Cosine similarity between sentences after PCA is also investigated, and is proved to be most effective.

6 Results

Figures 2, 4, 5, and 6 show the average F scores for tasks 1, 2, 3, and 4 respectively. For task 1, the system scores are shown alongside the "score" of the secondary assessor, who essentially performed this task (with the caveat that they did not judge sentences in the irrelevant documents). Within the margin of error of human disagreement, the assessor lines can be thought of as representing the best possible performance, and are fairly close to the scores for the second assessor last year.

Last year, the top systems were performing at the level of the second assessor, but this year there is a large gap between the second assessor and the systems. Moreover, nearly all systems had low average precision and high average recall. These two observations seem to imply that systems are much too lenient with what they accept as relevant or novel. Some runs with the lowest F scores actually achieved the highest precision of any run in task 1.

We cannot simply say that the difference in performance is due to the inclusion of irrelevant documents. In task 2, where systems are given all relevant sentences and therefore no interference from irrelevant documents, performance is much lower than in the same task last year. It may be that the systems have overly tuned to the 2003 data.

The systems all scored within a very small range, mostly between $0.36 - 0.4$ for relevant sentences and $0.18 - 0.21$ for novel. Precision is very uniform, but recall varies a lot. Last year, the best runs were also very close to one another; this year, the bottom systems have caught up, but the top systems have not improved very much.

Event topics proved to be easier than opinion topics. Figure 3 illustrates this for task 1, where every run did better on event topics than on opinions. The gap between opinions and events in task 1 is also larger than last year. The same gap exists in task 3, but in tasks 2 and 4, where all relevant sentences are provided, performance on opinion topics is much improved, and some runs do better on opinion topics than events. Thus, we can conclude that identifying sentences containing an opinion remains a hard problem.

Scores for task 2 (Figure 4) and task 4 (Figure 6) are shown against a baseline of returning all relevant sentences as novel. Most systems are doing better than this simplistic approach, both by F score and precision, indicating that the algorithms are successfully being somewhat selective.

It is also surprising how little the systems seem to benefit from training data. Overall scores did not improve between tasks 1 and 3, and from task 2 to task 4, novel sentence retrieval actually decreased significantly (see Figure 7). To be fair, this analysis needs to be balanced across groups, as tasks 3 and 4 had fewer runs and fewer groups participating, and some groups use radically different approaches in the presence of training data. But whereas last year additional training data helped relevant sentence retrieval markedly, this year there is no improvement.

7 Conclusion

This is the third and final year for the novelty track. We have examined a particular kind of novelty detection, that is, finding novel information within documents that the user is reading. This is by no means the only kind of novelty detection. Another important kind is detecting new *events*, which has been studied in the TDT evaluations. There, the user is monitoring a news stream and wants to know when something new, such as a plane crash, is first reported. Yet a third is the problem of returning new stories about a known topic, studied in the TREC filtering track and also in TDT topic tracking and story link detection.

We have seen here that filtering and learning approaches can be applied to detecting novel relevant information within documents, but that it remains a hard problem. Because the unit of interest is a sentence, there is not a lot of data in each unit on which to base the decision. Allowing arbitrary passages would make for a much more complicated evaluation.

The exploration into event and opinion topics has been an interesting and fruitful one. The opinions topics are quite different in this regard than other TREC topics. By mixing the two topic types within each task, we have seen that identifying opinions is hard, even with training data, while detecting new opinions (given relevance) seems analogous to detecting new information about an event.

One interesting footnote to the novelty track has been the use of the data outside the track. We know of two scenarios, namely summarization evaluation in DUC and an opinion detection pilot in AQUAINT, which have made use of topics from the novelty track. It's rewarding to see that this data is proving useful beyond the original scope of the track.

References

[1] 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.

[2] Ian Soboroff and Donna Harman. Overview of the TREC 2003 novelty track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, NIST Special Publication 500-255, Gaithersburg, MD, November 2003.

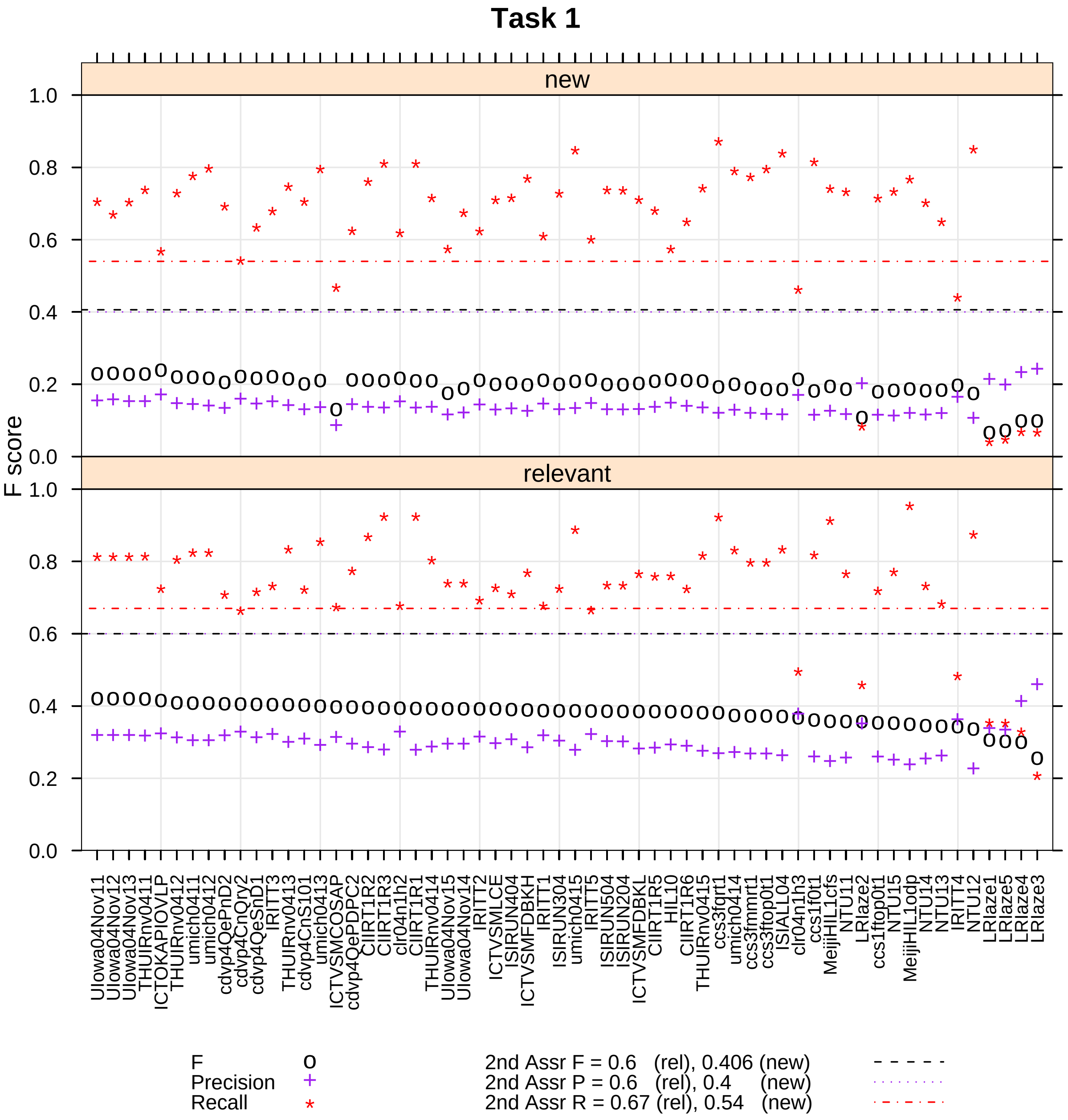

Figure 2: F, precision, and recall scores for Task 1, along with the "average score" of the secondary assessor. Runs are ordered by average F score for relevant sentence retrieval.

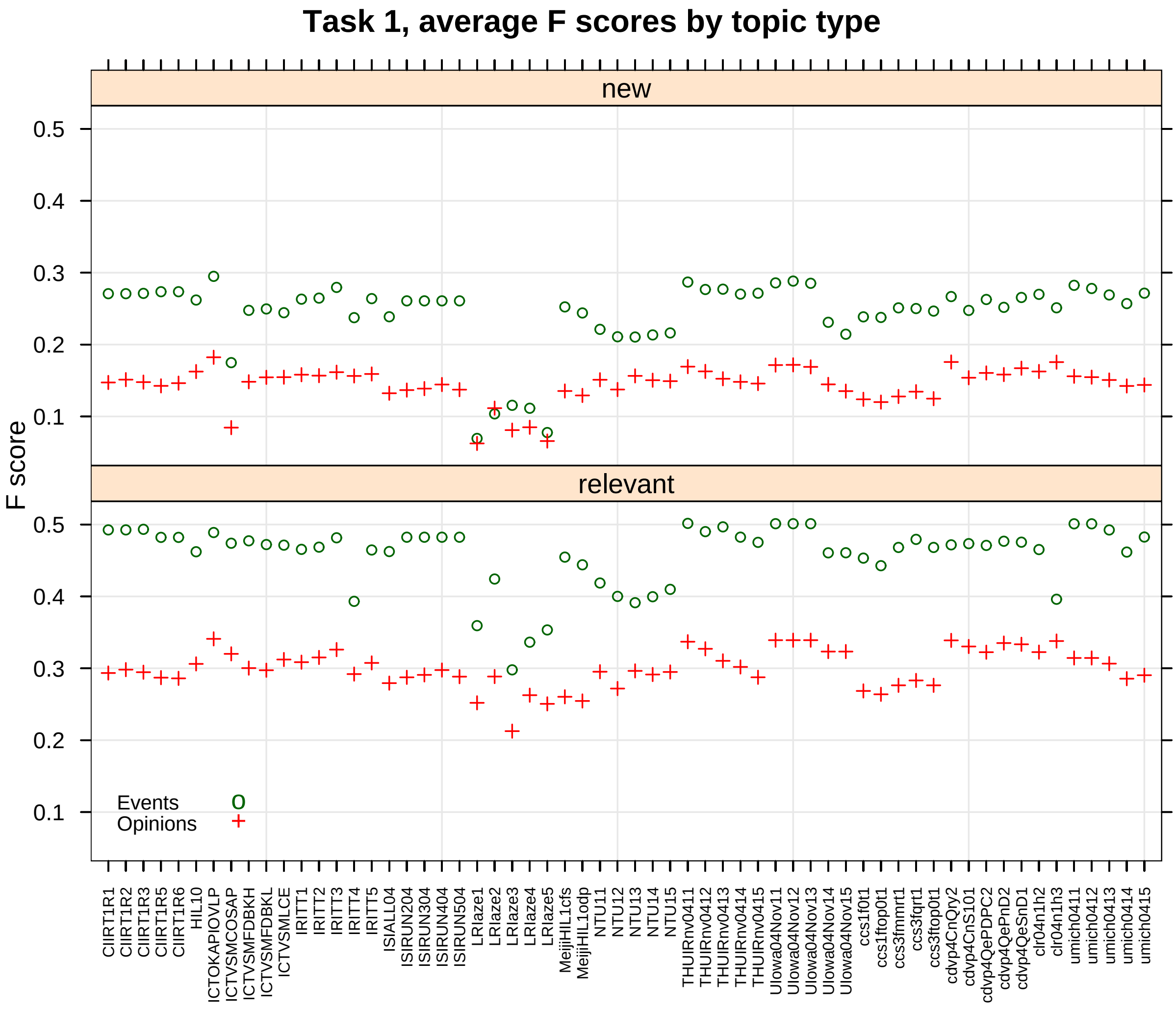

Figure 3: Average F scores per run for opinion and event topic types. Runs are grouped by tag for easier identification.

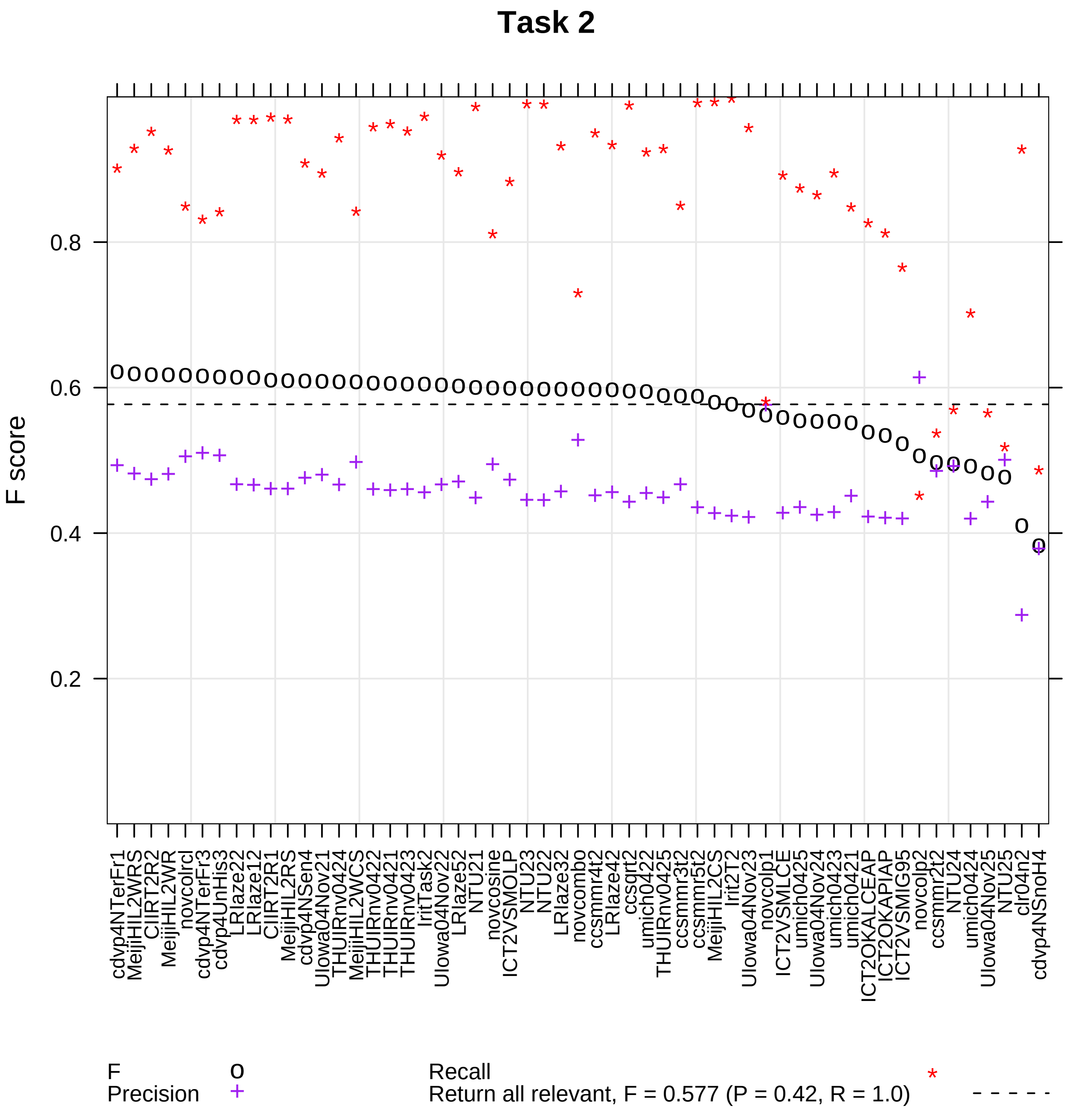

Figure 4: Scores for Task 2, against a baseline of returning all relevant sentences as novel.

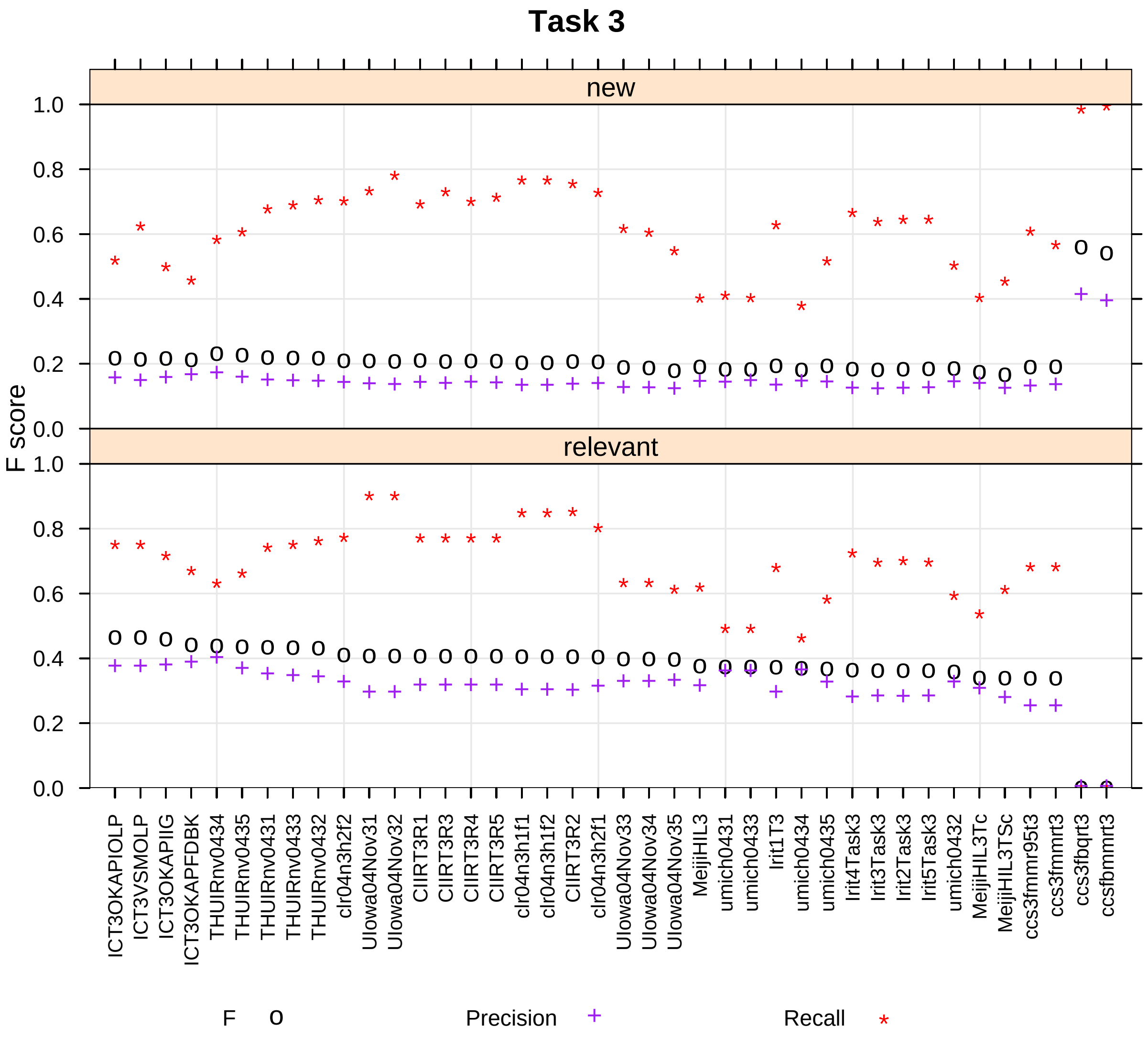

Figure 5: Scores for Task 3, ordered by average F score for relevant sentence retrieval.

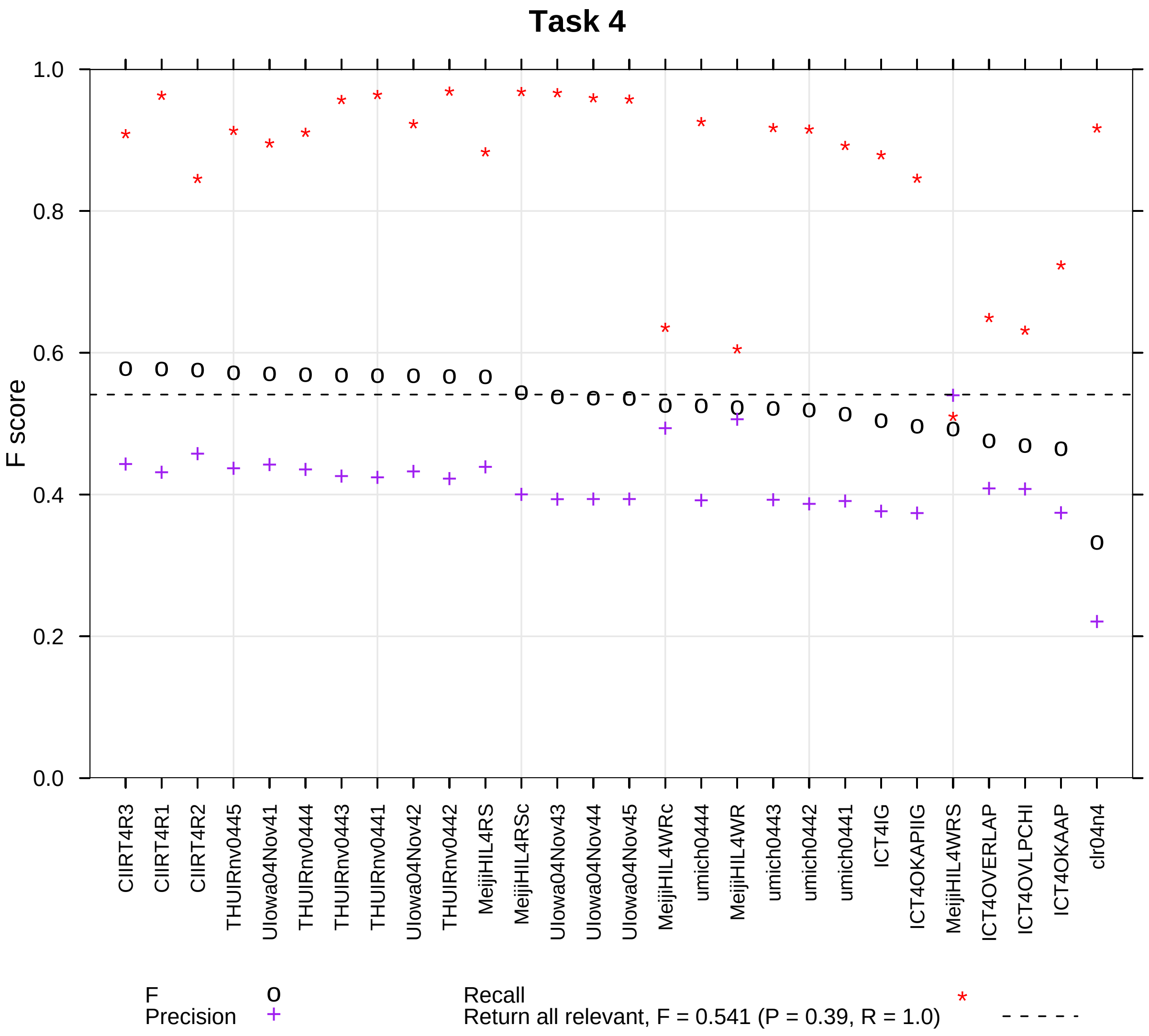

Figure 6: Scores for Task 4, against a baseline of returning all relevant sentences as novel.

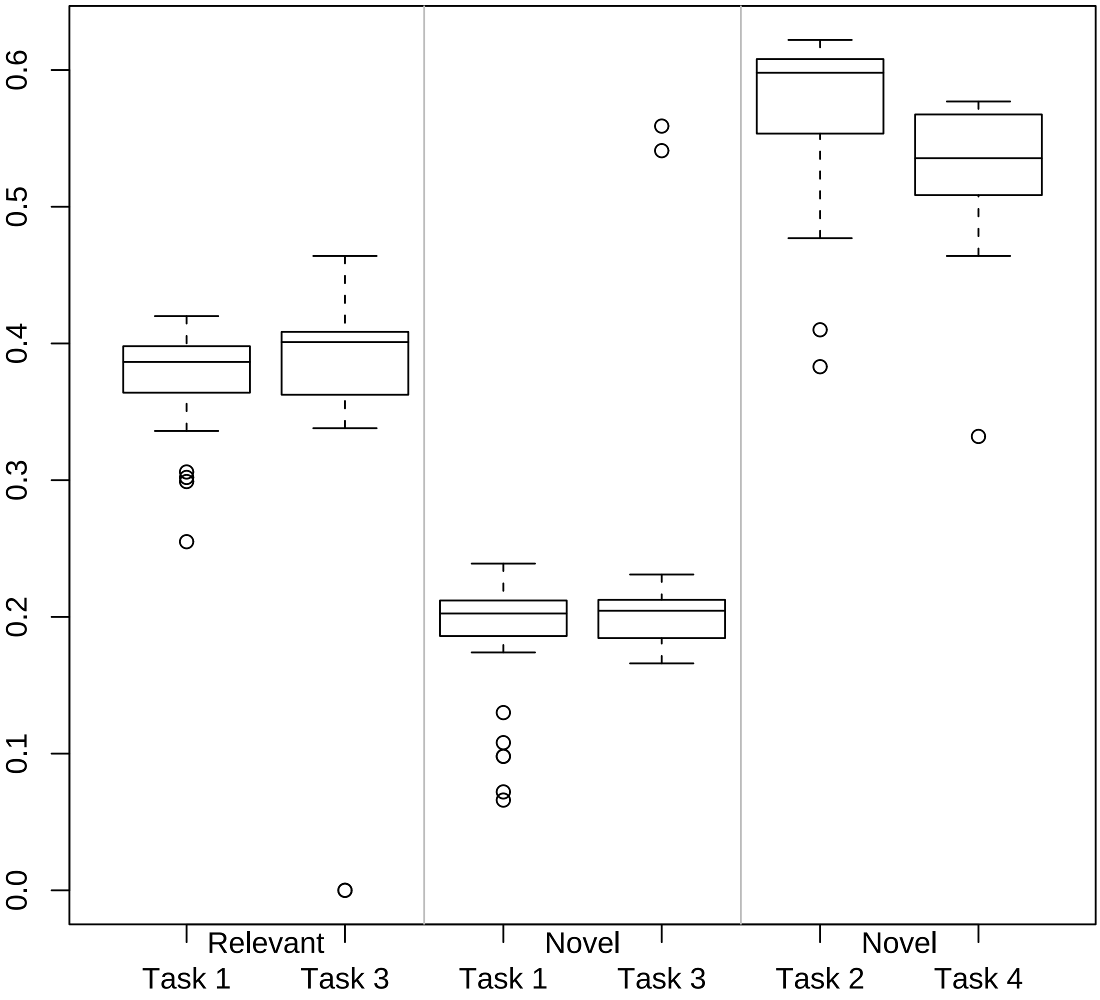

Figure 7: Comparison of F scores between Tasks 1 and 3, and between Tasks 2 and 4.