TREC 2003
Novelty Track Overview

Ian Soboroff
NIST

20 November 2003

Novelty Plenary Agenda

- Overview
  - Task overview
  - Data
  - Overall results
- Talks
  - Center for Computing Science/University of Maryland
  - Chinese Academy of Sciences – NLPR
  - Tsinghua University

The Novelty Task

- Given a topic statement and an ordered list of documents, identify sentences that are both *new* and *relevant*

- Four tasks

  1. Find all relevant and novel sentences. (same as last year)
  2. Given all relevant sentences, find all novel sentences.
  3. Given relevant and novel sentences from first five documents, find relevant and novel sentences in remaining documents.
  4. Given all relevant sentences, and novel sentences from first five documents, find novel sentences in remaining documents.

Participaction

		Runs submitted			
	Run prefix	Task 1	Task 2	Task 3	Task 4
Ctr for Computing Science / UMD	ccsum	5	4	3	5
Chinese Academy of Sciences (ICT)	ICT	5	5	5	5
Chinese Academy of Sciences (NLPR)	NLPR	5	5	5	5
CL Research	clr	4	1	5	1
Indian Institute of Tech. Bombay	IITB				1
IRIT	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			

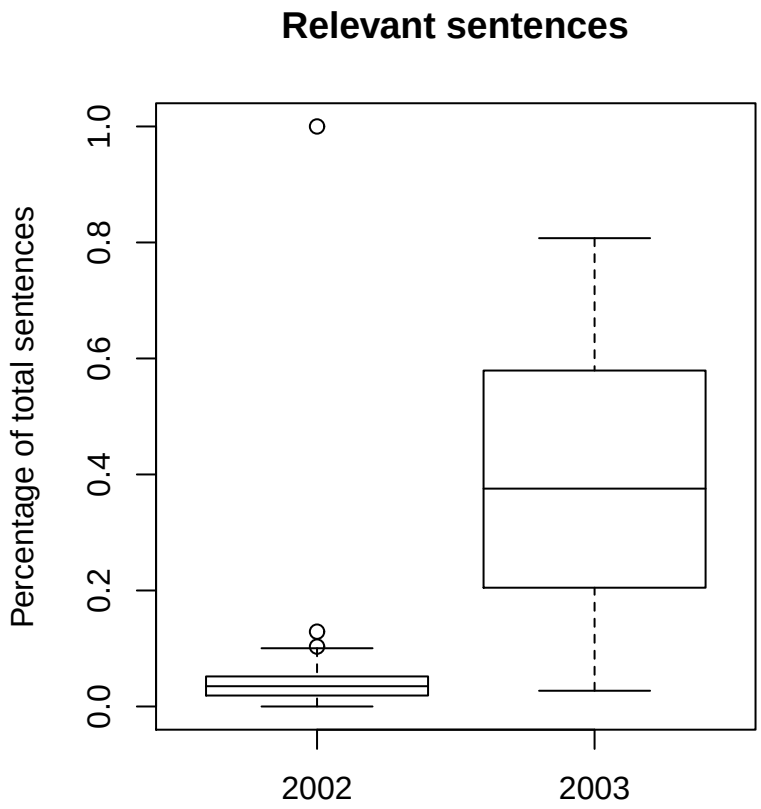
TREC 2002 Retrospective

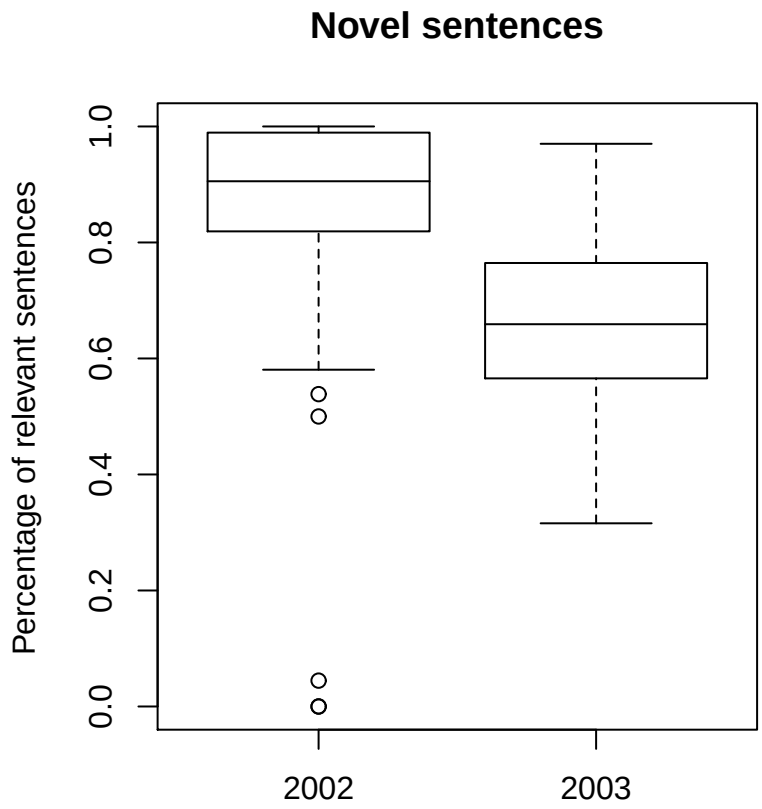
- Data issues dominated the task
  - Old topics led to assessor drift and disagreement
  - Very few relevant sentences (2%)
  - Almost all relevant sentences were novel (93%)
- Goals for 2003:
  - Avoid assessor drift by creating new topics
  - Try to encourage redundancy in the data
  - Allow participants to explore passage retrieval and/or filtering
  - Explore different topic types

2003 Novelty Data

- AQUAINT document collection
  - AP, NYT (6/1998–9/2000), Xinhua (1/1996–2/2000)
- 50 new topics
  - Events: about a particular news event
  - Opinions: about different points of view on a controversial subject
  - 25 relevant documents selected by topic author for each topic
  - 28 topics have documents from all three news sources
  - 21 topics have documents from two sources
- Documents ordered chronologically
- Segmented into sentences

Relevant and novel compared to last year





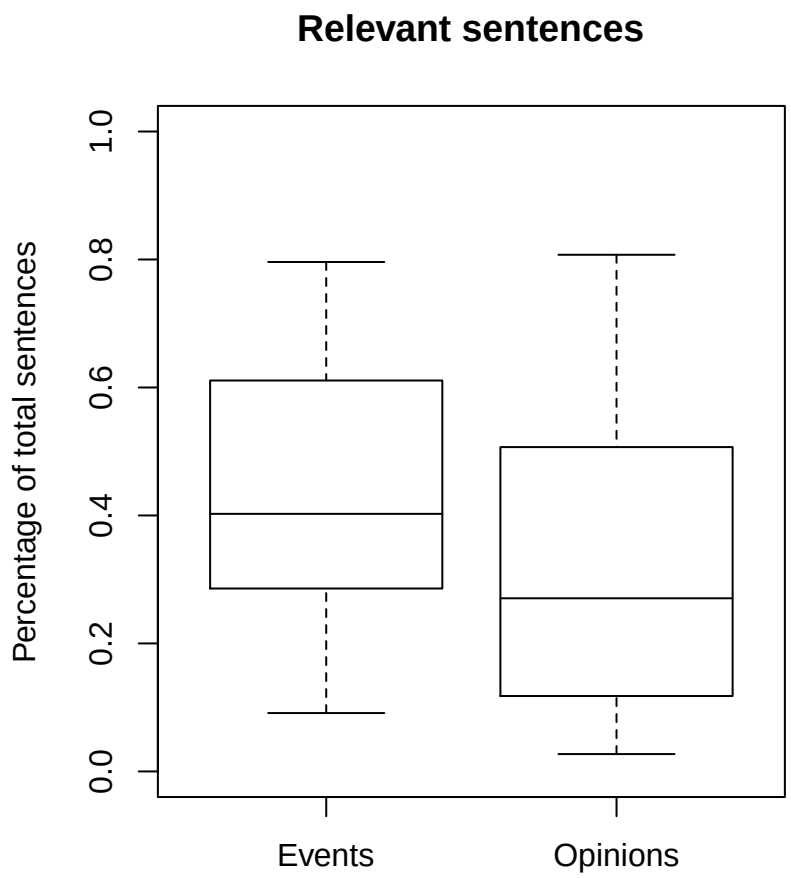
Assessor effects

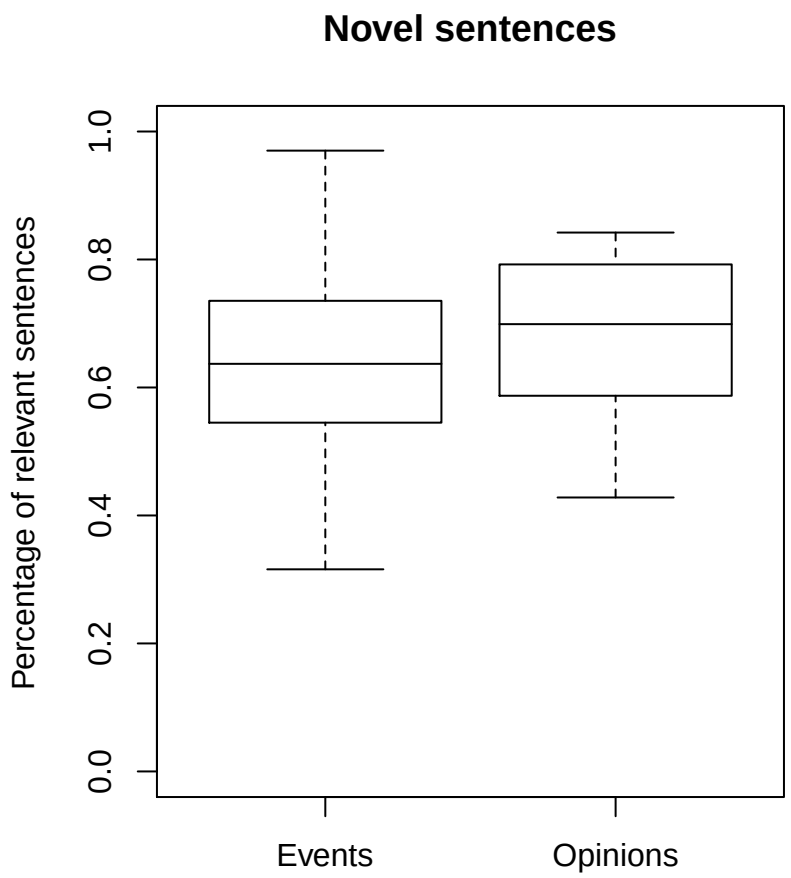
- Each topic was judged by two assessors
  - Primary assessor = topic author
  - Primary judgments are the official qrels
  - Secondary judgments used as a task baseline and to measure assessor disagreement

- Relevant sentences
  - Assessors significantly different from one another
  - Judged similarly whether as a primary or secondary assessor

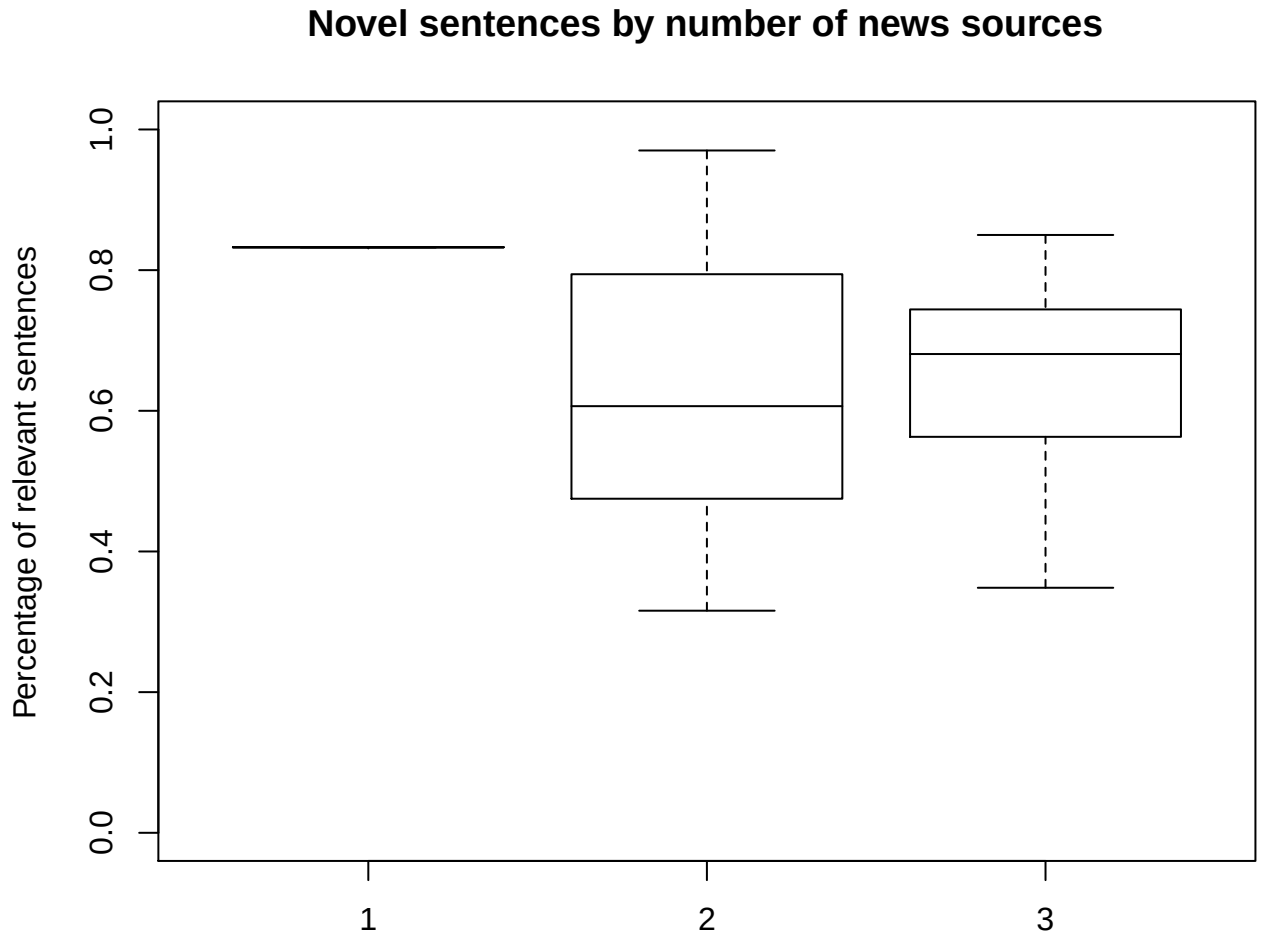
- Novel sentences
  - Less difference between assessors overall
  - Greater variation among primary assessors compared to secondary assessors

Relevant and novel by topic type





Did we increase the redundancy?

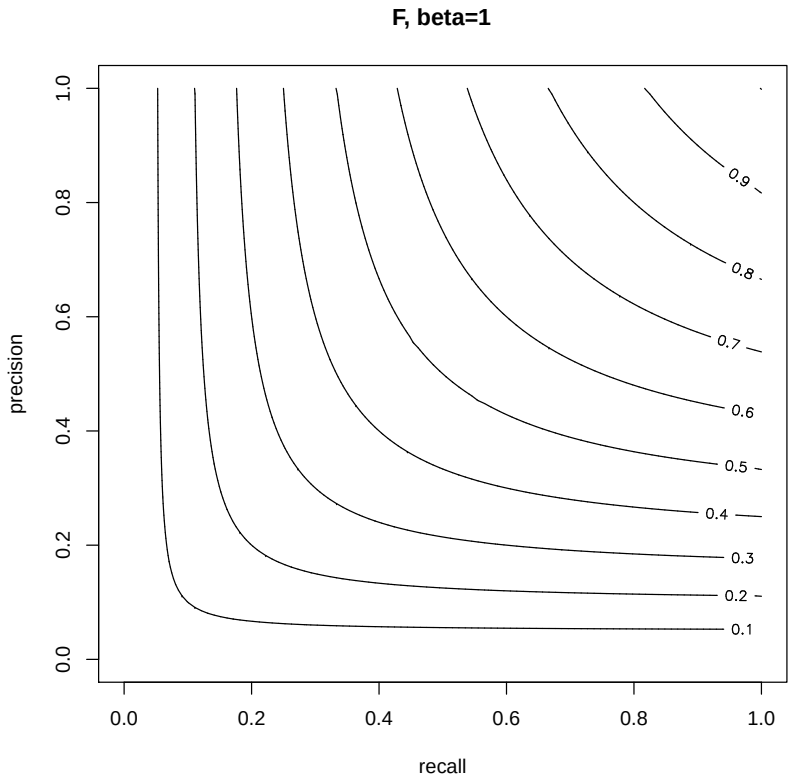


Evaluation

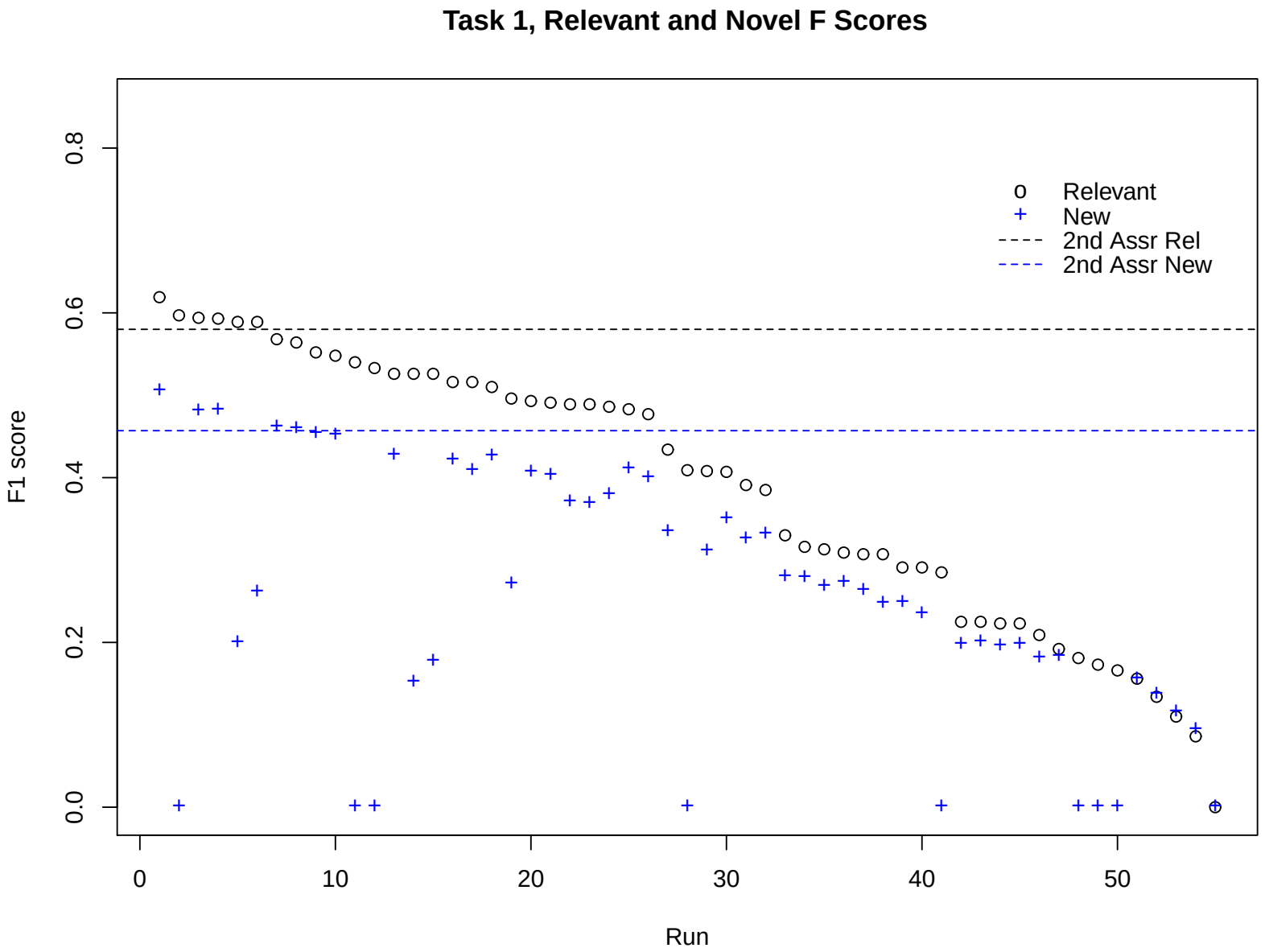
- Runs in task 1 and 3 scored for relevant sentence retrieval
- Runs in all tasks scored for novel sentence retrieval
- Measures
  - F ($\beta = 1$)

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

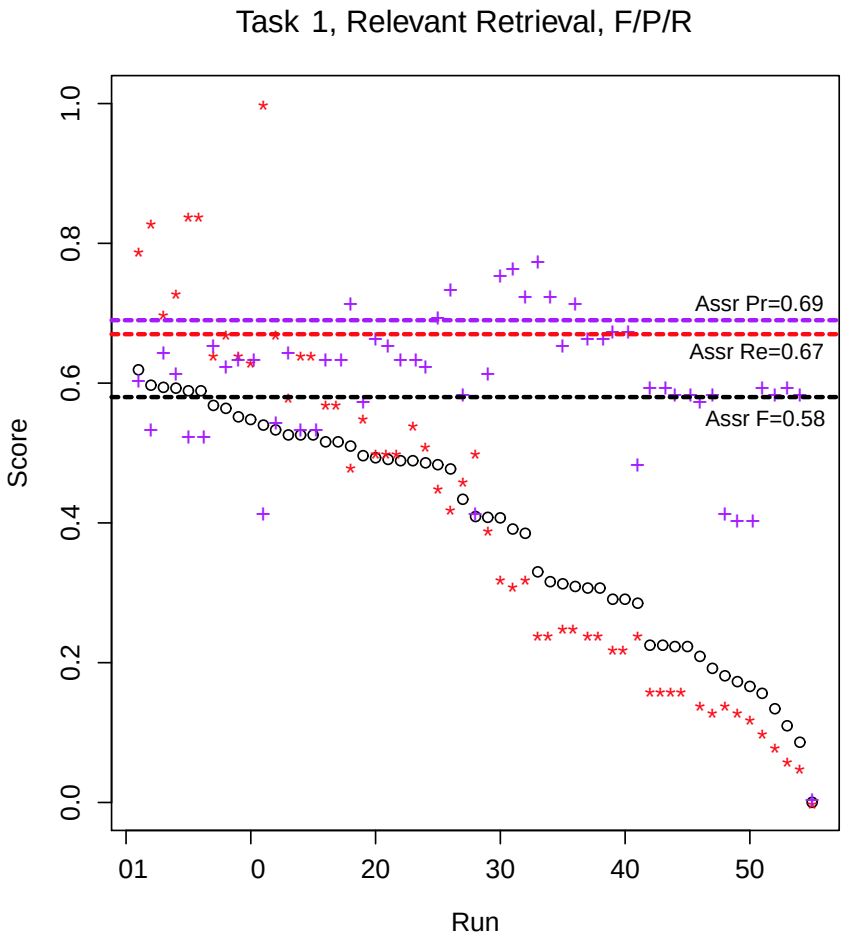
  - Set precision
  - Set recall

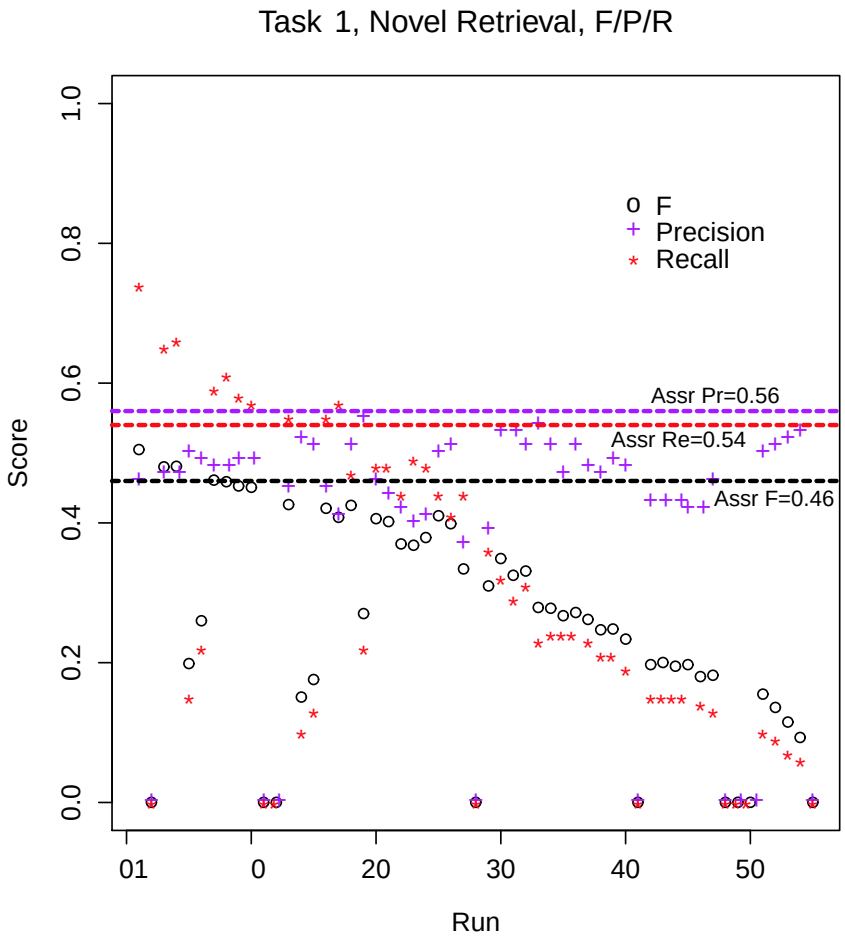


Task 1: Find all relevant and novel sentences

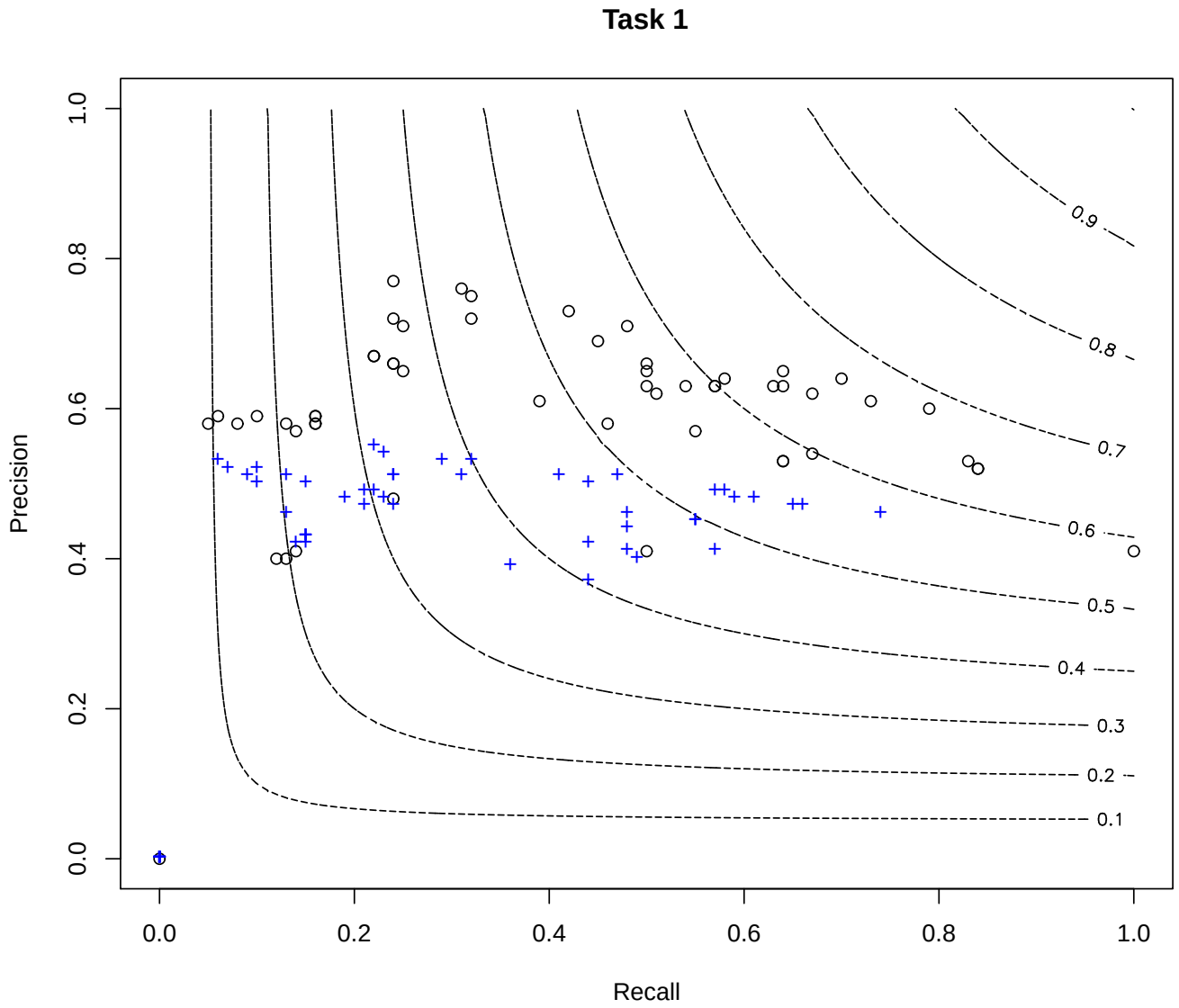


Task 1: F vs. Precision and Recall

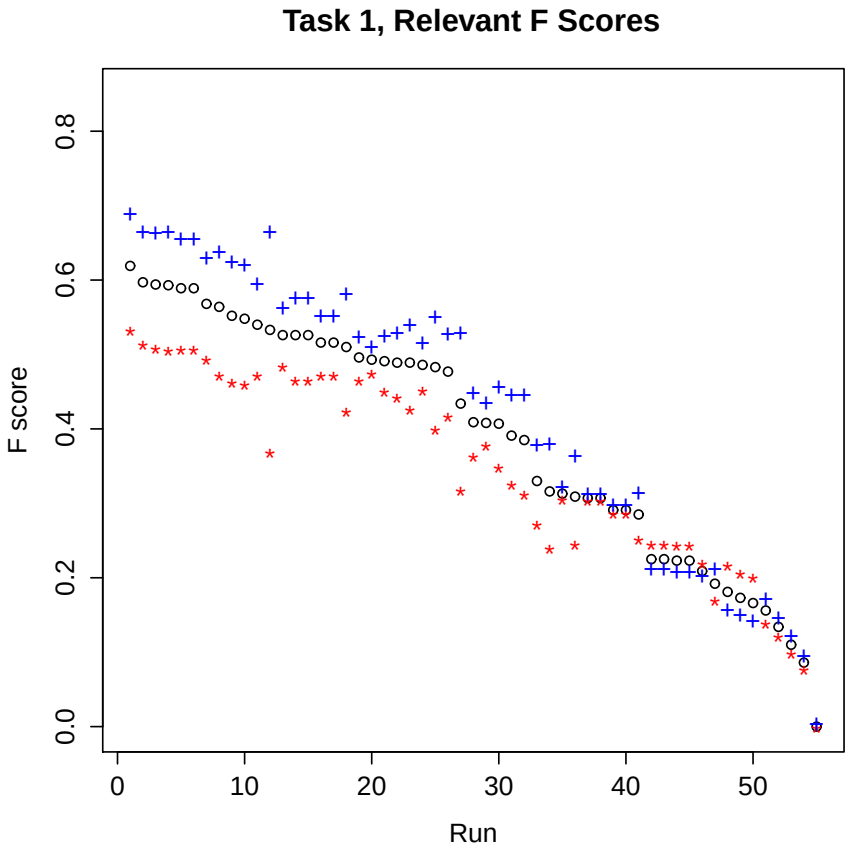


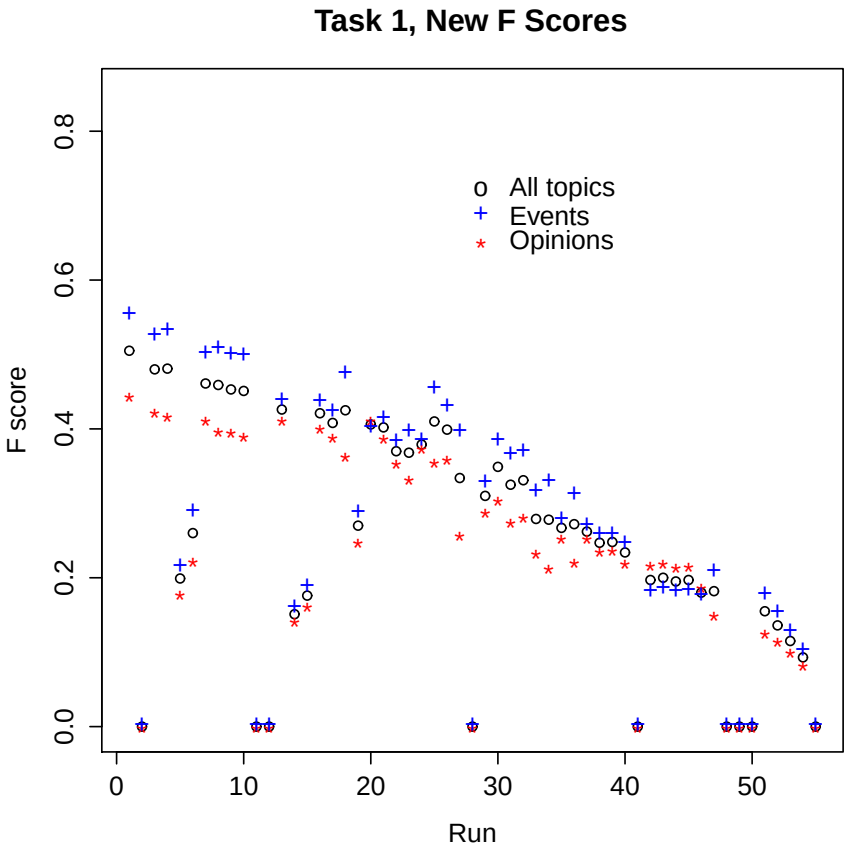


Task 1: F vs. Precision and Recall

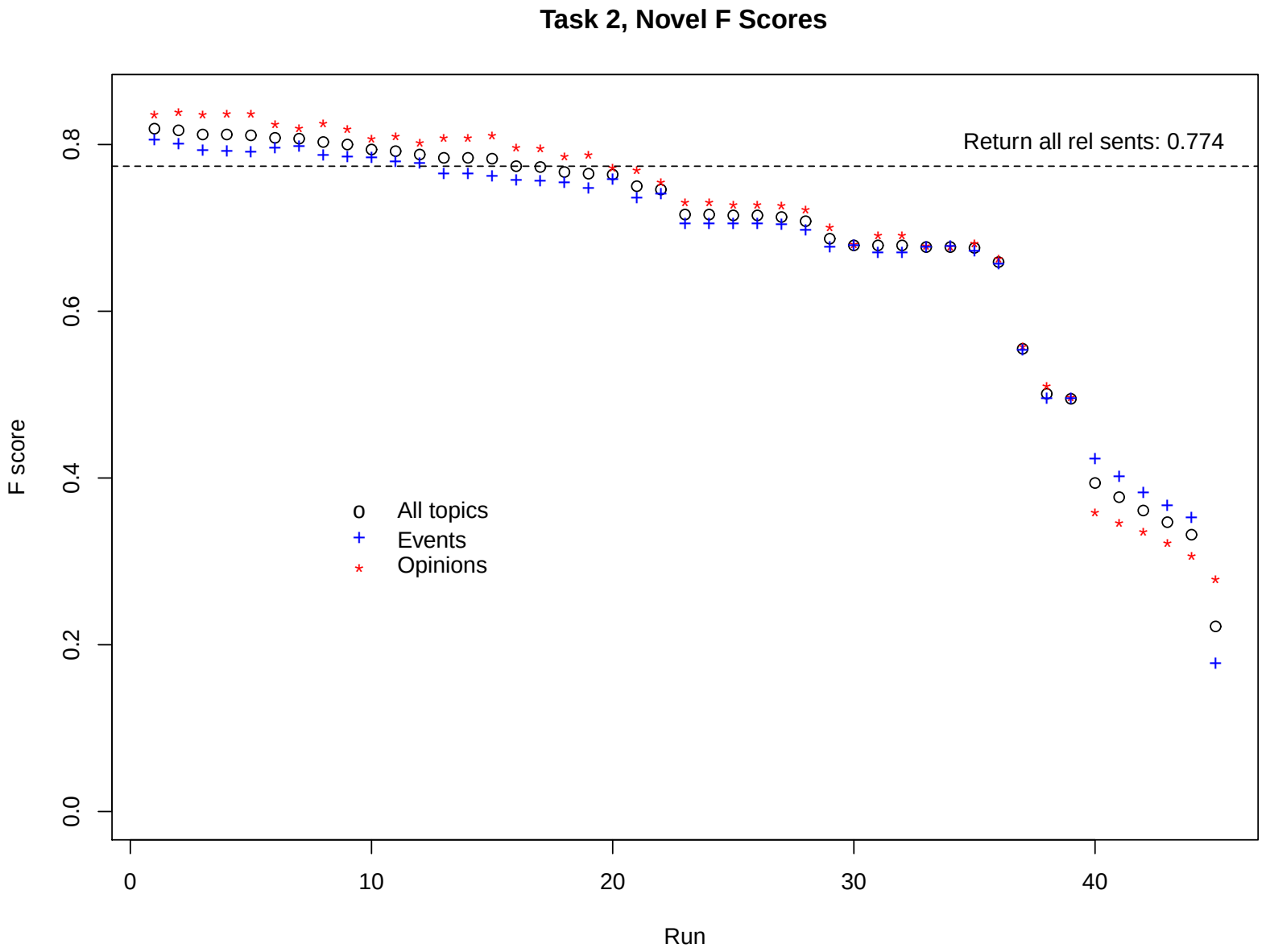


Task 1: Events and Opinions

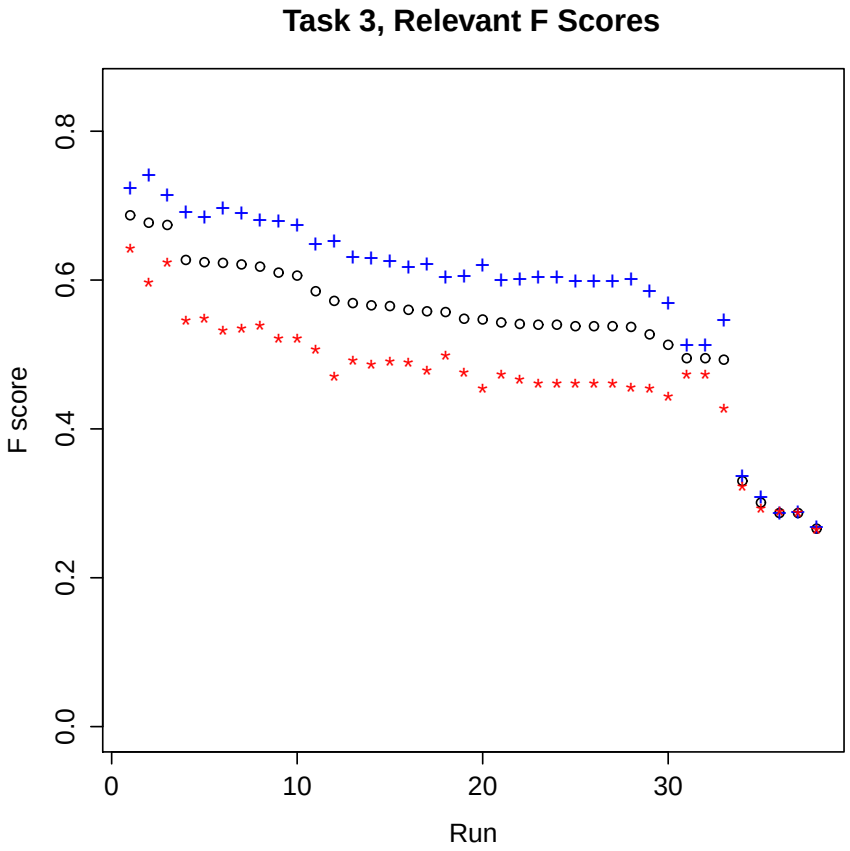


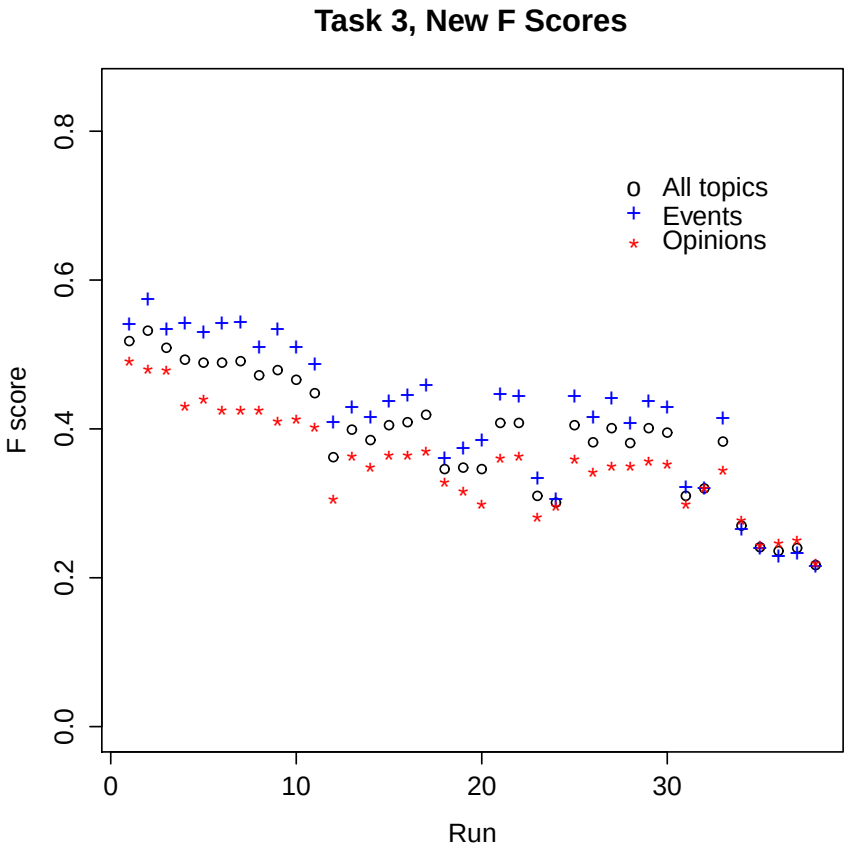


Task 2: Given all relevant, find all novel

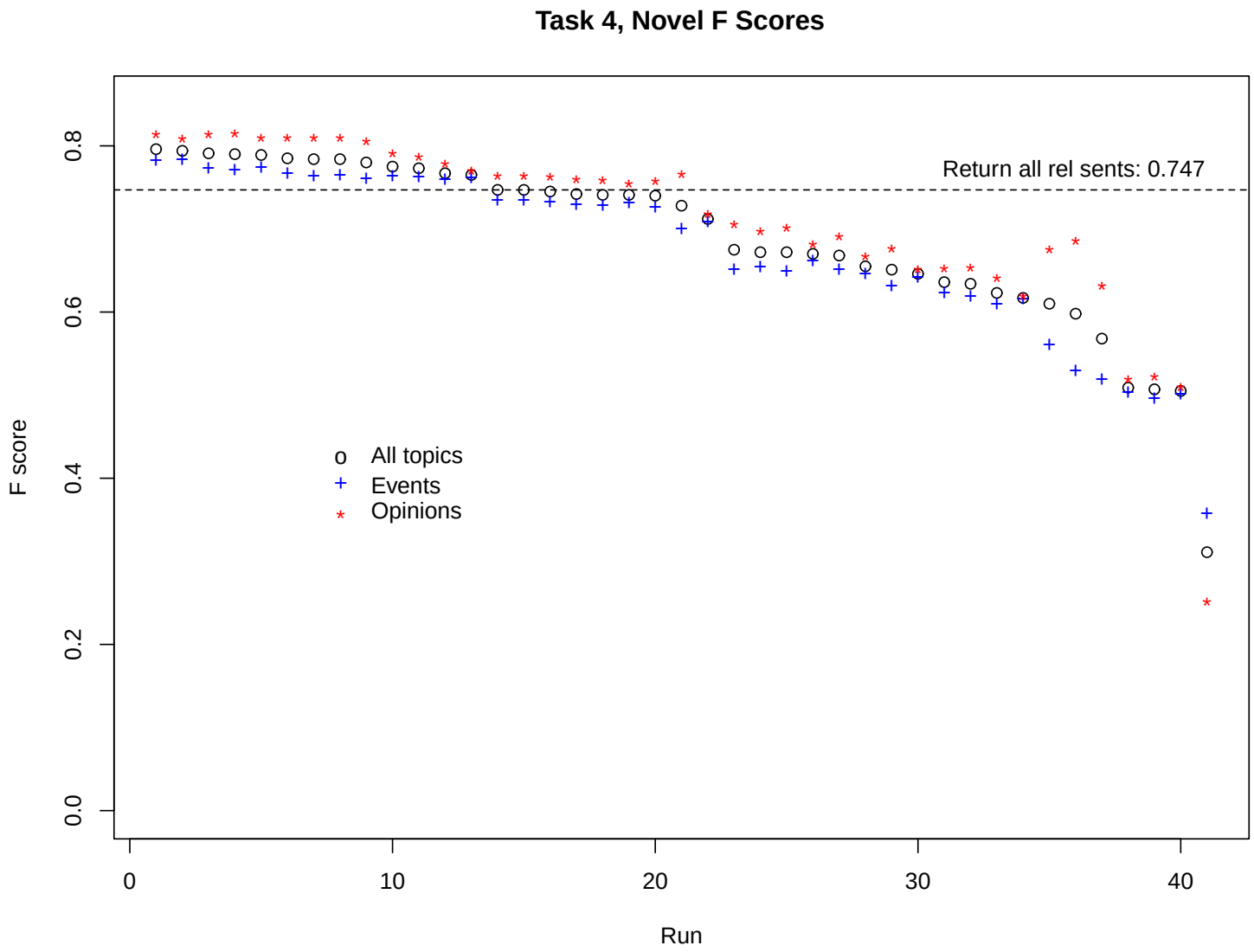


Task 3: Given relevant and novel in 5 docs, find the rest





Task 4: Given all relevant, novel in 5 docs, find remaining novel



Conclusions

- This year's novelty collection is better than last year's
- The best systems are performing at the level of a human
  - . . . but we don't know how much people disagree in this task
- Novelty is harder than relevance
- Training sentences help
  - Clear advantage in having sample relevant sentences
  - Novel training sentences were less useful
- Opinions are harder than events (without relevance training data)
- F measure is reflecting average recall more than precision