Определение новизны информации в новостном кластере

Определение новизны информации

- Определение новизны информации – важная и нерешённая задача.
- Проблема в общем виде:
  - поток информации и пользователь
  - в некоторый момент времени есть известная информация (известная пользователю)
  - Задача: извлечение новой информации из потока и предъявление пользователю

Конкретная задача

- Новостной кластер – набор документов по поводу некоторого события.
- Аннотация – краткое описание события, составленное из предложений документов кластера.
- В некоторый момент времени в кластер приходит ещё N документов.

Вопросы:

✔ **Что нового произошло?**

✔ **Как должна измениться аннотация?**

 Как новое отобразить в аннотации?

 Какие предложения аннотации должны быть заменены?

Конференция TREC

- Создана при поддержке Национального Института Стандартов и Технологий (NIST) и Департамента Защиты США.
- Проект был запущен в 1992 как часть программы TIPSTER Text.
- Назначение: поддержка исследований в области извлечения информации при помощи обеспечения инфраструктуры, необходимой для крупномасштабной оценки методов извлечения информации.

Постановка задачи «Определение новизны» в TREC

- **Данная задача разрабатывалась в TREC в 2002 – 2004 годах**

- **Постановка задачи: Дано упорядоченное множество документов, разделённое на предложения, и краткое описание(топик) к данному множеству.**

- **Задача: Найти важные(релевантные) и новые предложения.**

Постановка задачи-1

То есть по сути задача делится на две части:

1. Обнаружение значимых (важных) предложений.

   (identifying relevant sentences)

2. Выявление из этих значимых предложений, предложений несущих новую информацию.

   (novelty detection)

Постановка задачи-2

4 дисциплины:

- Task 1. Дан набор документов и топик, определить все релевантные и новые предложения.
- Task 2. Даны релевантные предложения во всех документах, определить все новые предложения.
- Task 3. Даны релевантные и новые предложения в первых 5 документах, найти все релевантные и новые предложения в остальных документах.
- Task 4. Даны релевантные предложения во всех документах и новые предложения в первых пяти, найти новые предложения в остальных документах.

Входные данные-1

- AQUAINT collection.
  - **New York Times News Service (Jun 1998 – Sep 2000),**
  - **AP (also Jun 1998 – Sep 2000),**
  - **Xinhua News Service (Jan 1996 – Sep 2000).**

  - Данная коллекция содержит сильную избыточность информации, и таким образом мы имеем меньше новой информации, повышая реализм задачи.

Входные данные-2

- Специалисты NIST сделали 50 кратких описаний новостей из данной коллекции.
- Новости были 2-ух типов: События (events) и Мнения (opinions).
- В описании топика содержался тег с его типом (участники заранее знали тип топика).
- Документы были хронологически упорядочены и разбиты на предложения.
- Предложения объединялись вместе, представляя собой единое множество документов к топику.

Оценка результатов-1

- Каждый топик был проанализирован двумя независимыми экспертами из NIST.
- Эксперты из набора документов выбрали релевантные предложения, и из этих предложений выбрали те, которые являются новыми.
- Некоторое преимущество экспертов перед системами, ввиду присутствия нерелевантных документов.

Оценка результатов-2

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
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

Оценка результатов-3

Введём следующие обозначения:

- M – число «правильных» предложений, то есть предложений, выбранных обоими экспертами и системой участником.
- A – число предложений выбранных экспертами.
- S – число предложений выбранных системой.

Оценка результатов-4

Тогда:

- **R = M / A** – эффективность поиска. (Recall)
- **P = M / S** – точность поиска. (Precision)

Проблемы:

I. R = 1 , P -> 0

II. P = 1 , R -> 0

=> Среднее значение R и P не является объективным критерием.

Оценка результатов-5

- Вариант решения: F-мера (F-measure)

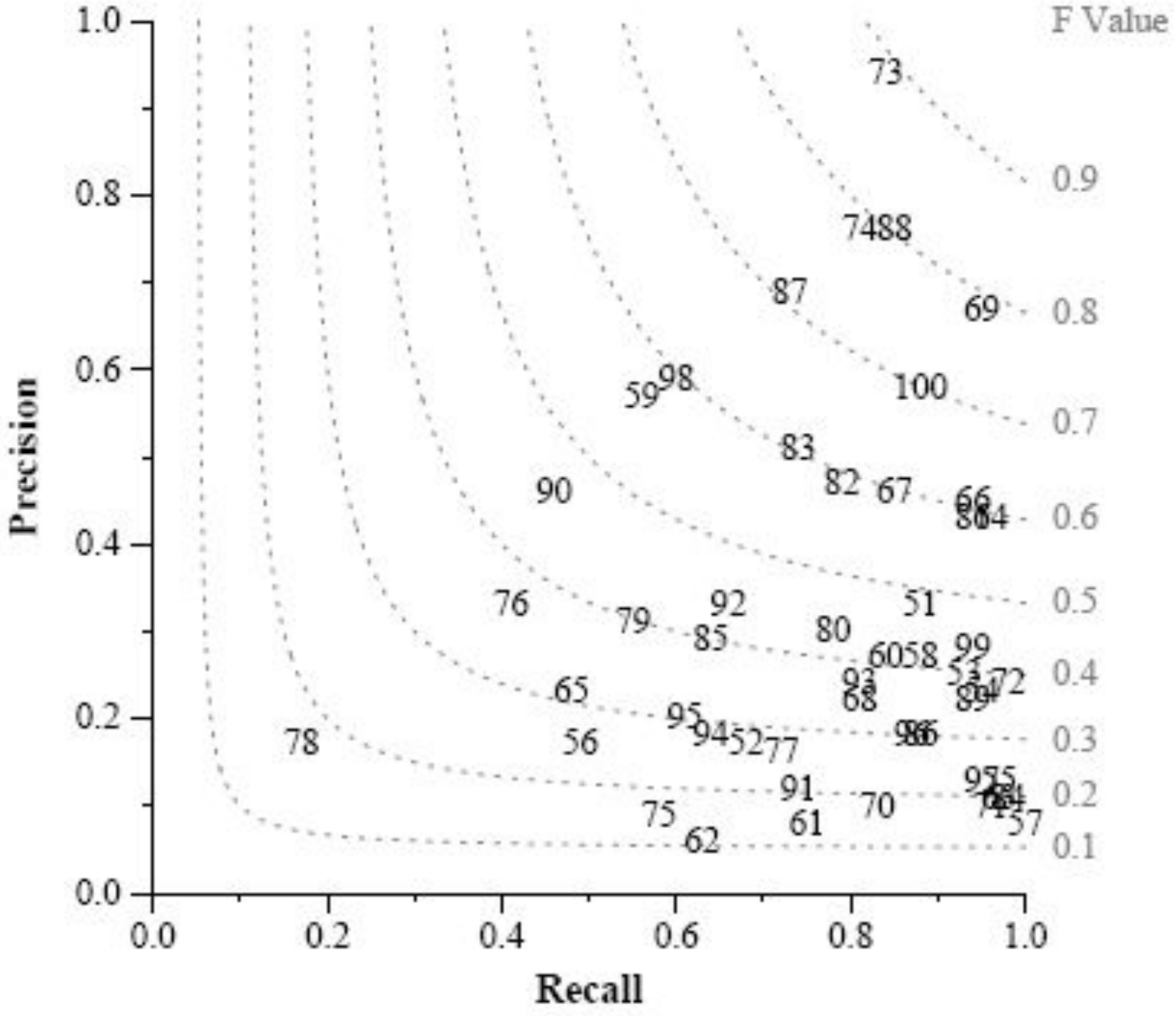
Общий вид:

$$F_\beta = \frac{(1+\beta^2)\cdot(\mathrm{precision}\cdot\mathrm{recall})}{(\beta^2\cdot\mathrm{precision}+\mathrm{recall})}$$

- F-measure, используемая на Novelty track:

$$F = \frac{2\cdot\mathrm{precision}\cdot\mathrm{recall}}{(\mathrm{precision}+\mathrm{recall})}.$$

Оценка результатов-6



Участники

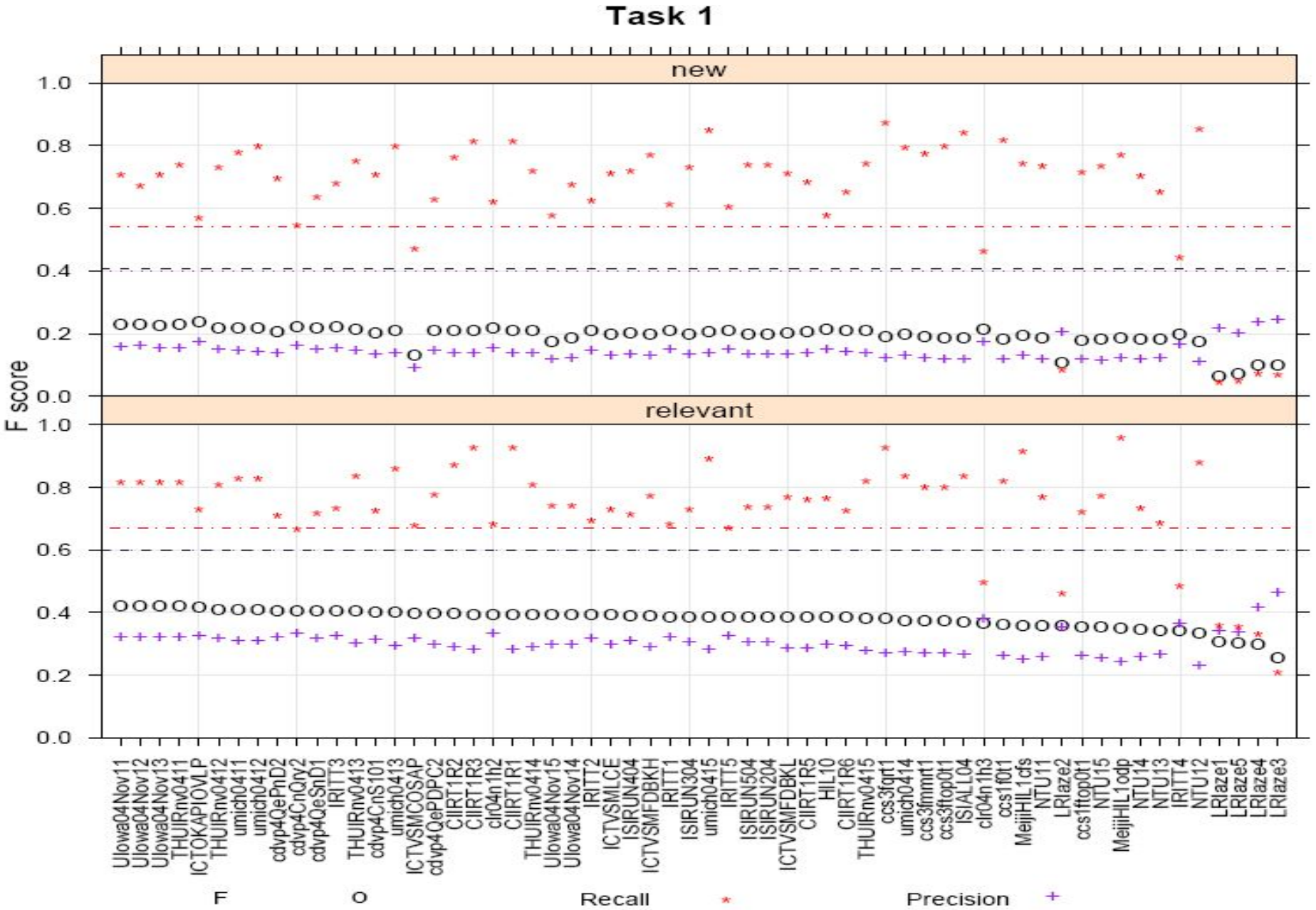
Table 2: Organizations participating in the TREC 2004 novelty track

		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			

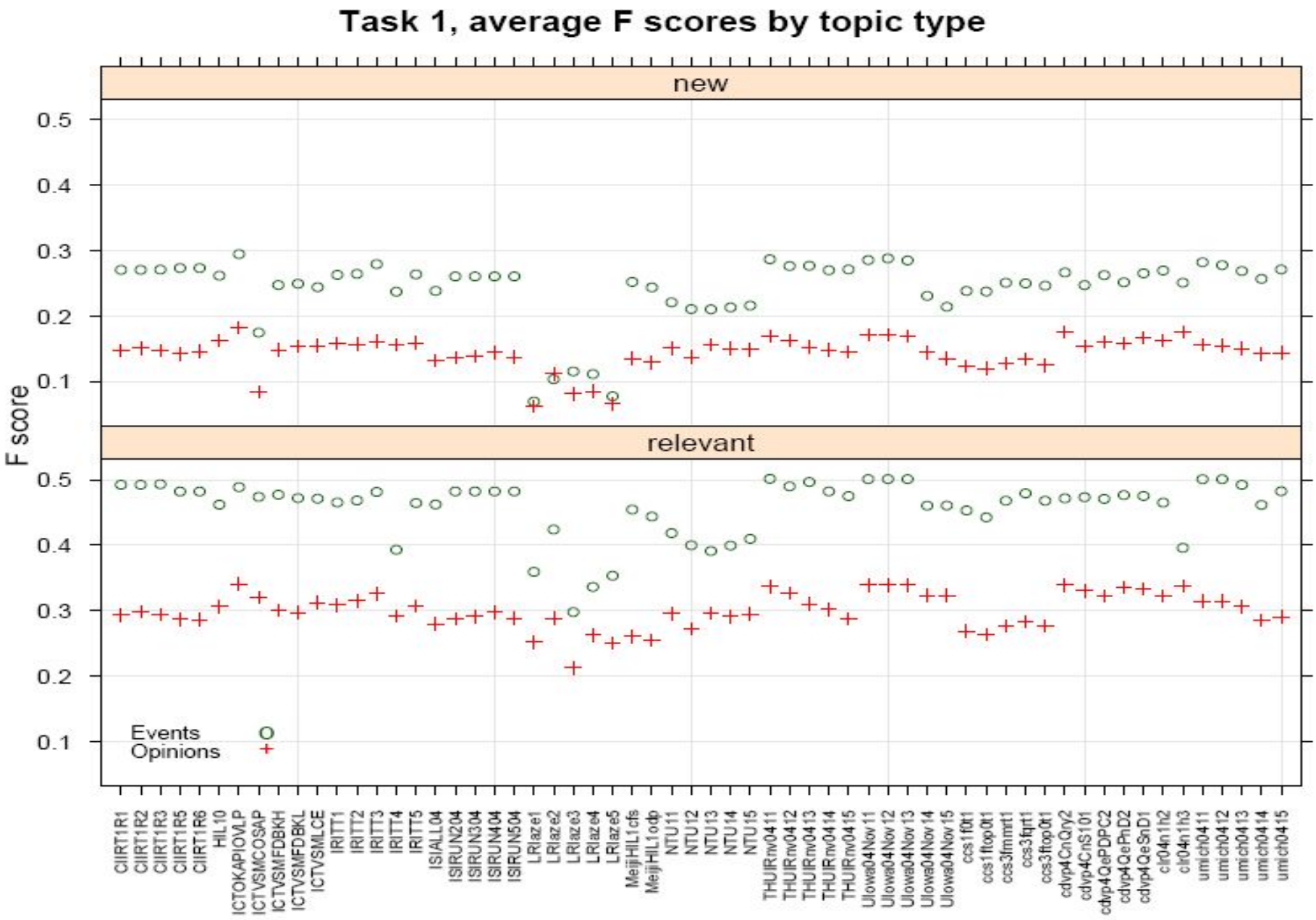
Результаты - 1

- В целом не очень высокие абсолютные результаты.
- Среднее значение F – меры:
  - 0.36 - 0.4 для задач обнаружения релевантных предложений.
  - 0.18 - 0.21 для задач обнаружения новой информации.
- Топики типа «Событие» оказались заметно проще топиков типа «Мнение».

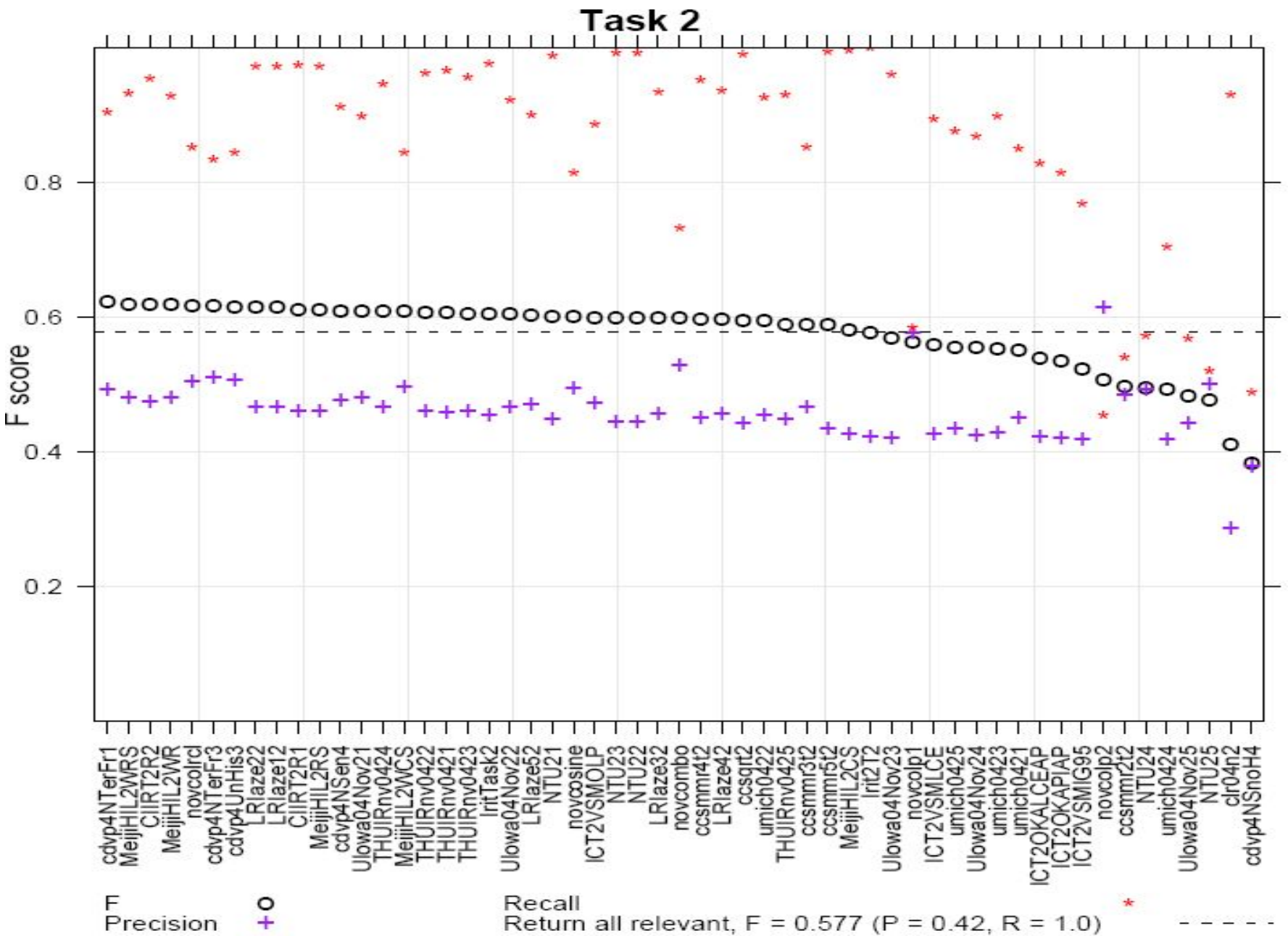
Результаты - 2



Результаты - 3



Результаты - 4



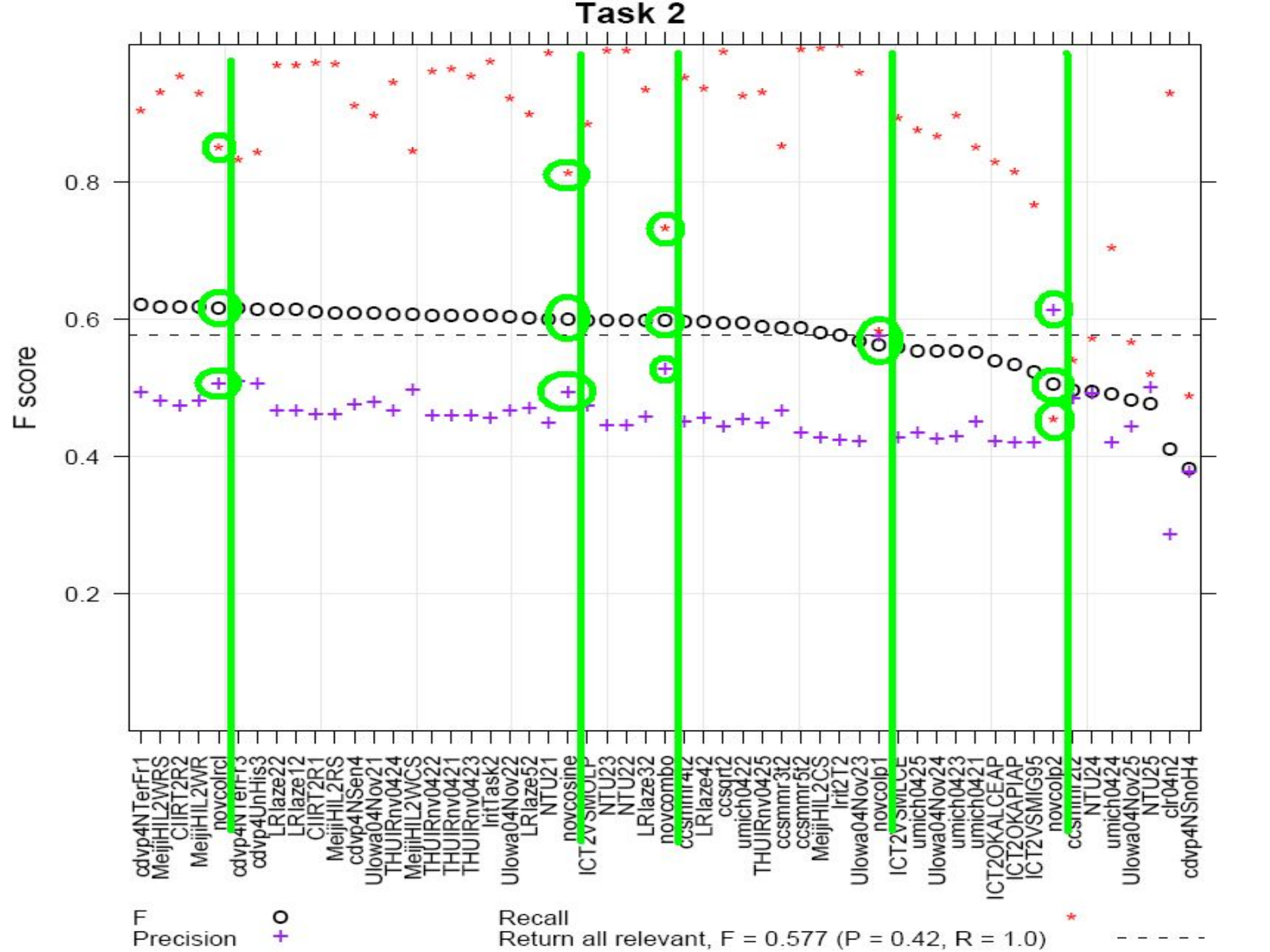
Анализ результатов TREC

Task 2. Даны релевантные предложения во всех документах, определить все новые предложения.

- Данная дисциплина ближе всего нашей задаче.

Колумбийский университет и система SumSeg:

- Основное направление – извлечение новой информации.
- Большое количество новых идей и подходов к решению задачи.
- Высокие результаты:


Task 2
F score
0.8
0.6
0.4
0.2
cdvp4NTerFr1
MeijiHIL2WRS
ClIRT2R2
MeijiHIL2WR
novcolrcl
cdvp4NTerFr3
cdvp4UnHis3
LRlaze22
LRlaze12
ClIRT2R1
MeijiHIL2RS
cdvp4NSen4
Ulowa04Nov21
THUIRnv0424
MeijiHIL2WCS
THUIRnv0422
THUIRnv0421
THUIRnv0423
IritTask2
Ulowa04Nov22
LRlaze52
NTU21
novcosine
ICT2VSMCLP
NTU23
NTU22
LRlaze32
novcombo
ccsmmr4t2
LRlaze42
ccsqrt2
umich0422
THUIRnv0425
ccsmmr3t2
ccsmmr5t2
MeijiHIL2CS
Irit2T2
Ulowa04Nov23
novcolp1
ICT2VSMLCE
umich0425
Ulowa04Nov24
umich0423
umich0421
ICT2OKALCEAP
ICT2OKAPIAP
ICT2VSMIG95
novcolp2
ccsmmr2t2
NTU24
umich0424
Ulowa04Nov25
NTU25
clr04n2
cdvp4NSnoH4
F
Precision
Recall
Return all relevant, F = 0.577 (P = 0.42, R = 1.0)

Особенности и основные идеи системы SumSeg-1

- Новая информация может появляться в сегментах больше или меньше одного предложения.
- Уход от прямого сравнения предложений на «похожесть».
- Новое слово – новая информация.
- Классификация предложений (работа с предложением в его контексте)
- Тщательная работа с местоимениями.

Особенности и основные идеи системы SumSeg-2

- Большое количество различных весов и порогов.
- База данных частотных характеристик слов.
- Анализ контекстных характеристик слов и корректировка весов с их учётом.
- Машинное обучение (автоматический подбор оптимальных коэффициентов, порогов и весов)
- Векторно - пространственная модель представления информации.

Векторно-пространственная модель-1

- Алгебраическая модель представления текстовых документов (в общем случае любых объектов) в виде вектора идентификаторов.
- Каждое пространство соответствует отдельному терму. Если терм встретился в документе, то его значение в векторе не равно нулю.
- Существует много методов по вычислению весов термов в векторе.
- Сравнения близости векторов по косинусу угла между ними:

$$\cos\theta = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\| \, \|\mathbf{v_2}\|}$$

Векторно-пространственная модель-2

- Пример: Пусть есть два предложения. «Мама мыла раму» и «Папа мыл автомобиль». Сравним предложения на «похожесть» при помощи ВПМ.

«Мама мыла раму»

папа	мама	мыть	автомобиль	рама
0	1/3	1/3	0	1/3

«Папа мыл автомобиль»

папа	мама	мыть	автомобиль	рама
1/3	0	1/3	1/3	0

$$COS = \frac{(0*1/3 + 0*1/3 + 1/3*1/3 + 0*1/3 + 0*1/3)}{2 * Sqrt((1/3)^2 + (1/3)^2 + (1/3)^2)} = 0.0962..$$

Направление дальнейшей работы

- Первоочередная задача – реализация векторно - пространственной модели и попытка её практического применения для обнаружения новой информации.
- Анализ весов и порогов, подбор оптимальных вариантов.
- Далее – анализ и реализация существующих и возможно создание новых методов и алгоритмов совершенствующих поиск (работа с различными частями речи, частотными характеристиками и т.д.)

The End