

Автоматическое аннотирование (реферирование)

Text Summarization

План

- Виды автоматического аннотирования
- Методы автоматического аннотирования
- Тестирование автоматического аннотирования

Автоматическое аннотирование (реферирование)


- Автоматическое аннотирование документа (совокупности близких по смыслу документов) - автоматическая технология, передающая в краткой форме основное содержание документа (совокупности документов)

- Аннотирование отдельного документа,
- Аннотирование совокупности документов - построение обзорного реферата

Назначение: быстрое ознакомление с содержанием документа (совокупности документов)

Примеры аннотаций: сниппеты

Поиск [Почта](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#) [Блоги](#) [Видео](#) [Картинки](#) [ещё](#)



Нашлось
29 тыс. ответов

☐ в найденном ☐ в Москве


Найти

расширенный поиск


[Мои нах](#)
[Настрой](#)
[Регион:](#)

[Разместить объявле](#)
[«автоматическое...»](#)
месяц


1

 [Автоматическое реферирование статей на русском языке / Хабрахабр](#)
5 мая 2011 Тема **автоматического реферирования/аннотирования** текста была поднята давно и было придумано множество способов ее реализации.
[habrahabr.ru](#) > [Писалось](#)


2

 [Автоматическое реферирование и аннотирование](#)
«Либретто» (разработчик — компания «МедиаЛингва»), обеспечивающую **автоматическое реферирование** и аннотирование русских и английских текстов (система встраивается в Word)
[do.gendocs.ru](#) > [docs/index-13506.html](#)


3

 [Автоматическое реферирование и аннотирование](#)
Автоматическое реферирование и аннотирование — одно из направлений компьютерной обработки естественно-языковых текстов.
[do.gendocs.ru](#) > [docs/index-208893.html](#)


4

 [автоматическое реферирование](#)
Большой англо-русский и русско-английский словарь. automated abstracting — Лингвистика: **автоматическое аннотирование, автоматическое реферирование** ...
[dic.academic.ru](#) > [dic.nsf/eng_rus...автоматическое](#)

5

 [Системы автоматического реферирования](#)
Системы **автоматического реферирования**. Искусство **реферирования**, или составления аннотаций, или кратких изложений материала, иными словами...
[rudocs.exdat.com](#) > [docs/index-34660.html](#)

6

 [Автоматическое реферирование](#)
Автоматическое реферирование (Automatic Text Summarization) - это составление коротких изложений материалов, аннотаций или дайджестов, т.е...
[bourabai.kz](#) > [dbt/internetica/autorefer.htm](#)

7

[Системы автоматического реферирования](#)
Системы **автоматического реферирования**. Удо Хан, Индерджит Мани, "Открытые Системы" #12/2000. В статье рассматриваются инструменты и методы **реферирования**...


Найти

☐ только в этом сюжете

[расширенный поиск](#)

Главные новости Мои новости Политика **Общество** Экономика В мире Спорт Происшествия Культура Наука Hi-Tech Интернет Авто

Семья Кристофа де Маржери вылетает в Москву, чтобы забрать его останки

МИР 24  [Семья Кристофа де Маржери вылетает в Москву, чтобы забрать его останки](#) 15:55

Родственники погибшего в авиакатастрофе в московском аэропорту "Внуково" главы французского нефтяного концерна Total [Кристофа де Маржери](#) вылетают в Москву, чтобы забрать его останки, сообщает ТАСС.

Интерфакс -
Россия



[Смертельное столкновение](#) статья 12:27 вчера

Примерно в полночь частный легкомоторный самолет Falcon с тремя членами экипажа и одним пассажиром на борту во время разгона столкнулся со снегоуборочной машиной.

ТАСС



[Кудрин: президент Total Кристоф де Маржери многое сделал для прихода инвестиций в РФ](#) 10:29 вчера

Самолет бизнес-авиации Dassault Falcon совершал рейс в Париж, на борту находились четыре человека - де Маржери, два пилота и стюардесса.



[Все видео](#)



[Все фото](#)


[Развернуть все сообщения](#)

Ещё по теме

[Расшифровка «черных ящиков» Falcon займет 2-3 дня](#) 14:50

[Путин о де Маржери: Россия потеряла «настоящего друга»](#) 18:02

Яндекс Директ

 [Продажа квартир в Говорово!](#)

Продажа квартир в Говорово! Свежие предложения с фото. Удобный поиск!

[Все квартиры](#) · [Все новостройки](#) · [Вся недвижимость](#)

[kvadroom.ru](#)



[Новостройки в Раменском, МО](#)

Квартиры от 1 650 000 руб. Юго-восток

Подмосковья, г. Раменское. 27 от МКАД

[Адрес и телефон gsestate.ru](#)



[Продажа квартир: ЖК «Москва А101»](#)

ЖК Москва А101: 3

Типы аннотаций

- Абстракты vs. Экстракты
 - Экстракты получаются извлечением фрагментов исходных текстов (обычно предложений) – основная применяемая технология
 - Абстракты порождаются – экспериментальные технологии
- Типы по основному содержанию
 - Индикативные – собственно содержание
 - Информативные – упор на цифры, данные
 - Оценочные – упор на мнения

Типы аннотаций по содержанию

- Индикативная:
- Авария произошла накануне на юго-востоке Москвы: экипаж отдела вневедомственной охраны двигался на служебном автомобиле по Волжскому бульвару и сбил молодого человека, который переходил дорогу с велосипедом на зеленый свет.
- Оценочная:
- **Ужасная по своей нелепости** трагедия произошла в доме на Митинской улице. Маленькая девочка утонула... в аквариуме!

Типы аннотаций-2

- **По фокусу**
 - Общее содержание
 - В ответ на запрос (Query-based - сниппет) – контекстная аннотация, тематически-ориентированная аннотация
- **По структуре**
 - Связная аннотация
 - Структурная аннотация
- **Одного документа или многих документов**
- **На том же языке, на другом языке**

Экстракты:

извлечение предложений

- Обработка документа – разделение на предложения
- Подсчет веса предложения на основе некоторых характеристик
- Упорядочение предложений по мере снижения веса
- Отбор предложений с максимальным весом для аннотации

Автоматическое аннотирование по многим документам (обзорное реферирование)

- Аннотации новостных кластеров
- Исторические справки
- Аналитические записки

Задачи:

- Отобразить основное содержание
- Отобразить новую информацию
- Снизить повторы
- Очень серьезные проблемы связности изложения

Пример аннотации (100 слов) по многим документам (Метод MMR – см. далее)

- **Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.**
- 1. Восьми улусам Якутии принадлежит 8 % акций компании.
- 2. Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад.
- 3. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.
- 4. Росимущество принадлежит 37 % акций АЛРОСА, минимуществу Якутии - 32 %, восьми улусам Якутии - 8 %, физическим и юридическим лицам - 23 %, из которых ВТБ владеет 10, 5 % акций.
- 5. Напомним, что до Александра Ничипорука АК АЛРОСА возглавлял Владимир Калитин (март 2002 - декабрь 2004 года), а еще ранее - действующий президент Якутии Вячеслав Штыров (1996-2002 годы).

Подходы к формированию экстрактов

- Частотные подходы:
 - самое важное должно быть частотно
- Подходы, основанные на графах
 - Центральность предложения
- Подходы, основанные на машинном обучении
 - Признаки предложения и их комбинирование
- Подходы, основанные на методах оптимизации

Частотные подходы

- Интуиция
 - Слова, которые часто повторяются в документе, имеют отношение к основной теме документа
 - Предложения, которые часто повторяются в разных документах, выражают основную тему документа
- Но также полезна информация из контрастной текстовой коллекции (tf.idf)

SumBasic

❑ Метод аннотирования, основанный на частотных характеристиках слов (Nenkova, Vanderwende, 2005)

❑ **Основная идея:** наиболее частотные слова исходного кластера с большей вероятностью должны оказаться в аннотации кластера:

n – число вхождений слова,

N – общее число токенов

$$p(w_i) = \frac{n}{N}$$

SumBasic-2

- ❑ Итеративный метод. На каждой итерации происходит расчет вероятностей слов, отбирается предложение с максимальной средней вероятностью слов:

$$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i \mid w_i \in S_j\}|}$$

- ❑ После отбора предложения происходит пересчет вероятностей для слов из отобранного предложения: вес слова резко снижается

Проблемы частотного подхода

- Разнообразие лексики
 - Правительство **Киргизии** передало для ратификации в законодательный орган... Парламент **Кыргызстана** в четверг примет окончательное решение о судьбе авиабазы США... стали главной причиной побудившей правительство **страны** принять такое решение

Обогащение пословного представления: лексические цепочки -1

❑ Использование информации об объектах и связях между ними, описанной в тезаурусах

❑ **Популярные подходы:**

➤ *Английский язык: **Wordnet***

Barzilay R., Elhadad M., 1999

➤ *Русский язык: **PyТез***

Лукашевич Н.В., Добров Б.В., 2009

Пример лексических цепочек на основе тезауруса RuТез

КОРТ	14
ТЕННИС	12
АВСТРИЙЦЫ	12
АВСТРИЯ	6
КИПРИОТЫ	16
КИПР	11
ХОРВАТЫ	10
СЕТ	6
ИГРОВАЯ ПАРТИЯ	5
ЧЕТВЕРТЬФИНАЛ	10
ПОЛУФИНАЛ	29
ПОЛУФИНАЛИСТ	2

МАТЧ	12
СПОРТИВНЫЙ ФИНАЛ	36
СПОРТИВНОЕ	54
СОРЕБНОВАНИЕ	8
СПОРТ	2
СПОРТСМЕН	1
ФИНАЛИСТ	
ЮЖНЫЙ, МИХАИЛ	23
РОССИЯНЕ	12
РОССИЙСКАЯ	10
ФЕДЕРАЦИЯ	6
ТЕННИСИСТ	
ЗАГРЕБ	70
ХОРВАТИЯ	36

Частотные методы: использование TF.IDF

- Частотное слово в исходном документе (наборе документов) может быть недостаточно тематическим,
- оно всегда частотно в данной коллекции
- Поэтому:
- Tf – в исходном документе
- Idf – в корпусе, из которого извлечен этот документ

Подходы, основанные на методах оптимизации

- Локальная оптимизация
 - Оптимизация выбора следующего предложения
- Глобальная оптимизация
 - Оптимизация некоторой функции (например, присутствие наиболее частотных слов или выражений в аннотации)
 - Линейное программирование

Maximal Marginal Relevance (MMR)

□ (Carbonell, Goldstein, 1998)

- Итеративный метод
- На каждой итерации производится ранжирование предложений-кандидатов
- В итоговую аннотацию отбирается одно с самым высоким рангом
 - ✓ Максимизировать сходство с исходным документов (набором документов)
 - ✓ Минимизировать сходство с уже отобранными в аннотацию предложениями
- При использовании в контекстно-зависимой аннотации – максимизируется сходство с запросом

Maximal Marginal Relevance (MMR)

Пусть:

Q – запрос к системе

S – множество предложений кандидатов

s – рассматриваемое предложение
кандидат

E – множество выбранных предложений

Тогда:

$$\text{MMR} = \arg \max_{s \in S} \left[\lambda \cdot \text{Sim}_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} \text{Sim}_2(s, s_j) \right]$$

Графовые методы автоматического аннотирования

- Вершины – предложения
- Дуги – сходство между предложениями
 - Косинусная мера
 - (Radev et.al., 2004) LexRank –
аннотирование многих документов

Косинусная мера сходства между предложениями новостного кластера

- Предложение ID dXsY означает Y предложение в X документе

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

Граф связей между предложениями

- Жирные линии соответствуют высокому весу

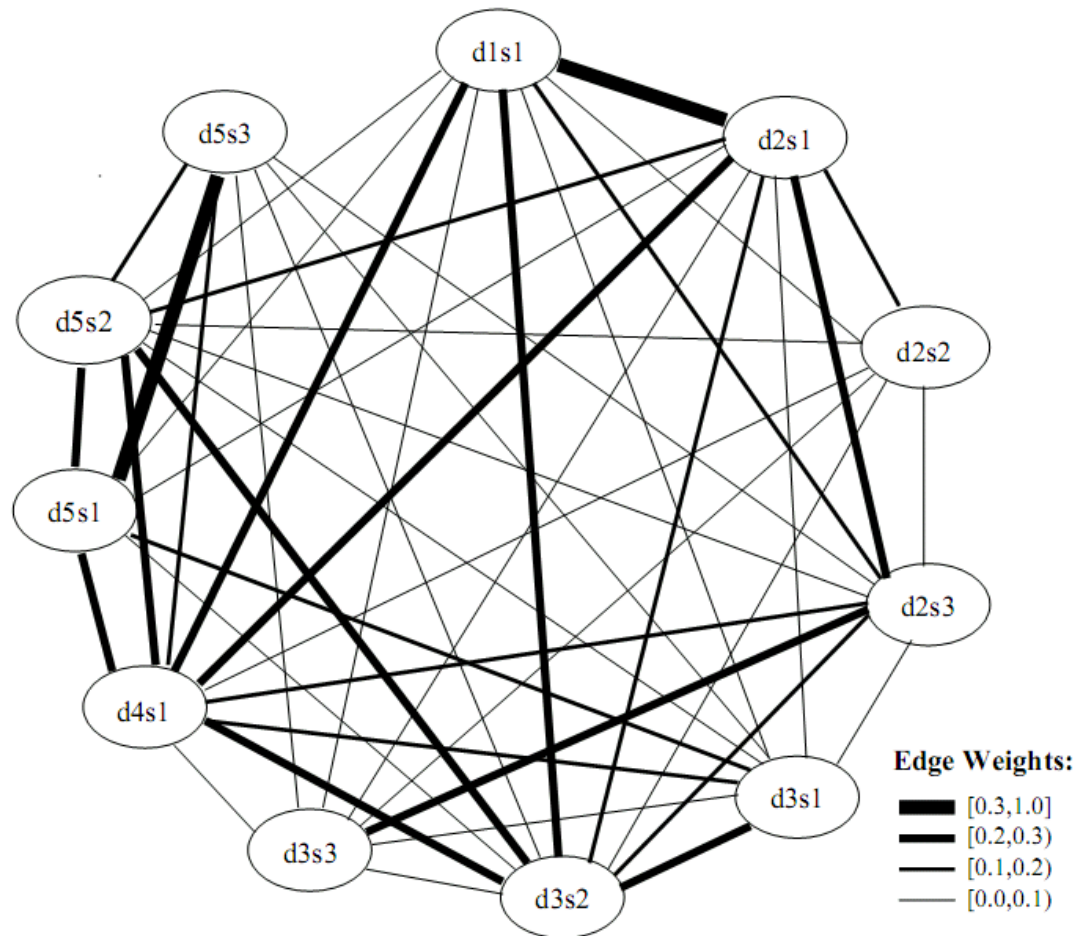


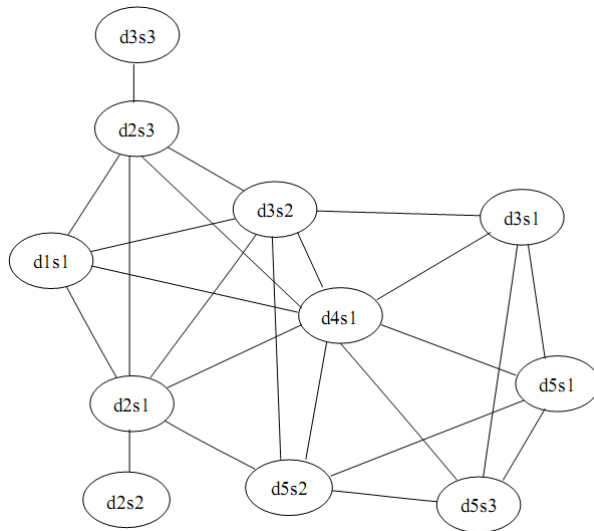
Figure 2: Weighted cosine similarity graph for the cluster in Figure 1.

Центральность предложения

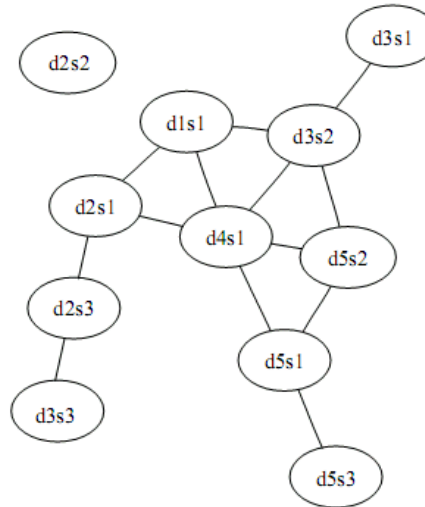
- В кластере похожих документов многие предложения в некоторой степени похожи друг на друга
- Интересны значительные степени сходства, определяемые заданным порогом (0.1, 0.2, 0.3). Тогда малые величины в матрице обнуляются
- Центральность предложения – предложение при заданном порого связано с максимальным числом других предложений. Это предложение и берется в аннотацию

ID	Degree (0.1)	Degree (0.2)	Degree (0.3)
d1s1	5	4	2
d2s1	7	4	2
d2s2	2	1	1
d2s3	6	3	1
d3s1	5	2	1
d3s2	7	5	1
d3s3	2	2	1
<u>d4s1</u>	9	6	1
d5s1	5	4	2
d5s2	6	4	1
d5s3	5	2	2

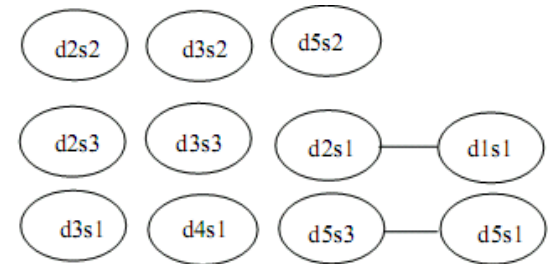
Граф сходства при разных порогах



0.1



0.2



0.3

- Важен выбор порога

Методы на основе разных
характеристик и машинного
обучения

Экспериментальная комбинация (Edmundson, 1969)

Вклад четырех признаков:

заголовки,
частотные слова,
ключевые слова,
позиция

$$Weight(S) = \alpha.Title(S) + \beta.Cue(S) + \gamma.Keyword(S) + \delta.Position(S)$$

Параметры были подобраны на обучающей выборке

Характеристики для извлечения предложений

- Частотные слова текста
 - Слова получают веса,
 - вес предложения зависит от весов слов
- Учет заголовка
 - Вес предложения увеличивается, если в нем присутствуют слова заголовка
- Присутствие ключевых слов и конструкций
 - Специальный список: *подчеркнем, основным результатом*
- Позиция в тексте
 - Предложения в начале более важны
- Связность с предыдущим
- Новизна информации

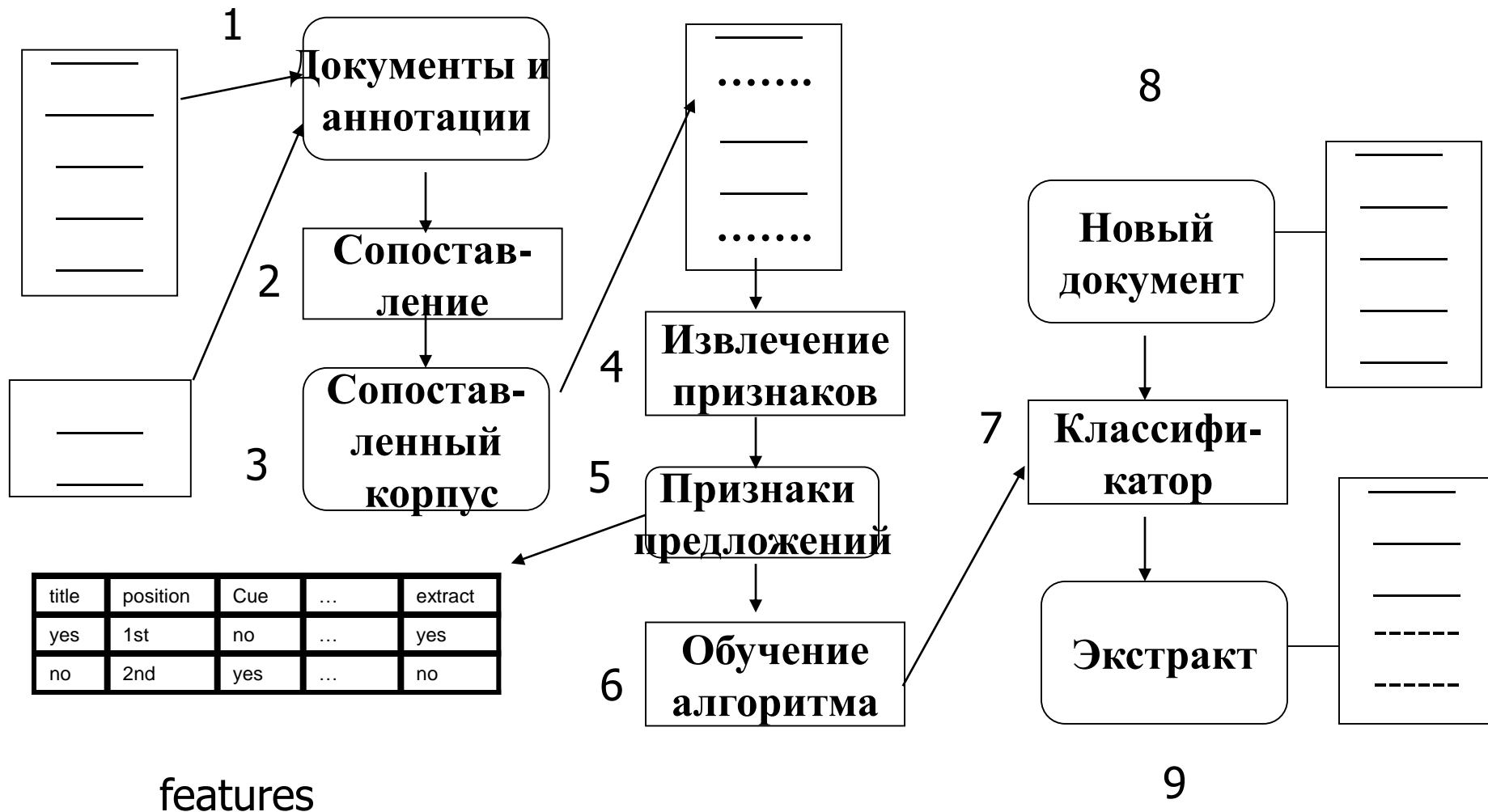
Комбинация на основе статистического подхода

- Метод из работы (Kupiec&al'95)
- Нужен корпус документов и экстрактов
 - Профессиональные абстракты

Сопоставление

- Отождествление похожих предложений в абстрактах и документах
- Ручная проверка

Обучение для извлечения предложений



Статистическая комбинация (признаки)

- Длина предложения (true/false)

$$\textit{len}(S) > u_l$$

- Ключевые фразы (true/false)

$$(S_i \cap \textit{DIC}_{cue}) \neq \phi$$

or

$$\textit{heading}(S_{i-1}) \wedge (S_{i-1} \cap \textit{DIC}_{headings}) \neq \phi$$

Признаки-2

- Позиция (discrete)
 - paragraph # $\{1, 2, \dots, 10\} \vee \{last, last-1, \dots, last-4\}$
 - in paragraph $\{initial, middle, final\}$
- Тематические слова
(true/false)
$$rank(S) > u_k$$
- Собственные имена
(true/false)

Метод Байеса для комбинирования признаков

Вероятность, что предложение Принадлежит экстракту

Теорема Байеса

Признаки в экстрактах

Вероятность предложения в экстракте

$$p(s \in E | f_1, \dots, f_n) = \frac{p(f_1, \dots, f_n | s \in E) \cdot p(s \in E)}{p(f_1, \dots, f_n)}$$

Признаки в корпусе

The diagram illustrates the Bayesian formula for feature combination. It features the equation $p(s \in E | f_1, \dots, f_n) = \frac{p(f_1, \dots, f_n | s \in E) \cdot p(s \in E)}{p(f_1, \dots, f_n)}$ centered on the page. Four annotations with arrows point to specific parts of the equation: 'Вероятность, что предложение Принадлежит экстракту' points to the left side of the equation; 'Теорема Байеса' points to the entire equation; 'Признаки в экстрактах' points to the numerator's first term $p(f_1, \dots, f_n | s \in E)$; and 'Вероятность предложения в экстракте' points to the numerator's second term $p(s \in E)$. Additionally, 'Признаки в корпусе' points to the denominator $p(f_1, \dots, f_n)$.

Оценка параметров

Предполагается
независимость

$$p(f_1, \dots, f_n \mid s \in E) = \prod p(\cancel{f_i} \mid s \in E)$$

$$p(f_1, \dots, f_n) = \prod p(f_i)$$

Оценка
подсчетом



$$p(s \in E)$$

Статистическая комбинация

- Результаты для отдельных признаков
 - позиция
 - ключевые слова
 - длина предложения
 - тематические слова
 - собственные имена
- Лучшая комбинация
 - Позиция+ключевые слова+длина
 - Проблемы методов, основанных на машинном обучении??

Проблемы аннотирования, основанного на извлечении предложений

- «лишние» обороты в предложении
- отсутствие связности изложения – местоимения

Four adults and one child died in the crash, which witnesses said occurred about 5 p.m., when it was raining, Albuquerque police Sgt. R.C. Porter said.

It aborted its first attempt and was coming in for a second try when it crashed, he said.

Отсутствие целостности в изложении

- Исходный текст:
- Supermarket A announced a big profit for the third quarter of the year. The directory studies the creation of new jobs. Meanwhile, B's supermarket sales drop by 10% last month. The company is studying closing down some of its stores.
- Экстракт:
Supermarket A announced a big profit for the third quarter of the year. The company is studying closing down some of its stores.

Автоматическое аннотирование: методы тестирования

- **Методы тестирования:**
 - **Внутреннее (intrinsic)**
 - Является ли текст связным и целостным
 - Содержит ли он главные темы документа?
 - Сравнение аннотации с идеальными аннотациями
 - **Внешнее (extrinsic)**
 - может ли аннотация использоваться вместо документа

Конференция Summac (1998)

- Extrinsic evaluation
- Оценка аннотирования посредством ручного рубрицирования
- Участникам рассылается 1000 текстов
- Темы текстов делятся на 5 подразделов внутри двух больших тем:
 - Мировая экономика
 - Экология
- Процедура:
 - Аннотации двух видов: 10% и лучшей длины
 - Эксперт по тексту должен отнести его к одной из десяти рубрик
 - Несколько экспертов рассматривают один и тот же текст
 - Среди текстов есть полные тексты и начала текстов

SUMMAC: подсчет результатов

Мировая экономика

экспорт в промышленности

внешняя торговля

международная борьба с наркотиками

иностранные производители автомобилей

налоги

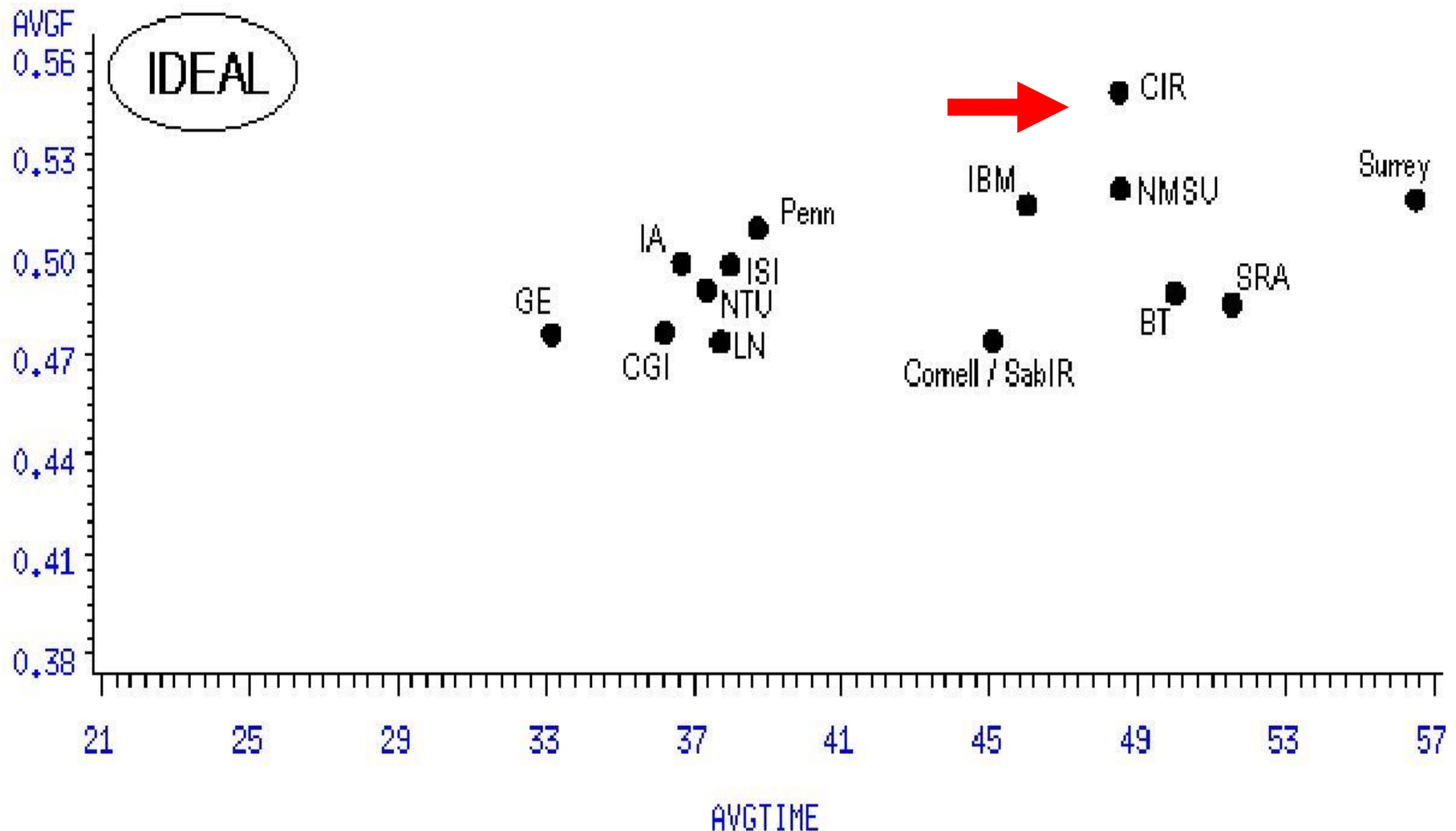
$$\text{Точность} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Полнота} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Fscore} = \frac{2 \bullet \text{Точность} \bullet \text{Полнота}}{(\text{Точность} + \text{Полнота})}$$

SUMMAC 1998 (NIST DARPA TIPSTER III)

Categ: F-Score vs. Time by Party for Best-Length Summaries



Оценка качества аннотаций

choosing sentences

N	Human	System
1	+	+
2	-	+
n	-	-

- Точность

$$\frac{TP}{TP + FP}$$

- Полнота

$$\frac{TP}{TP + FN}$$

contingency table

True Positive		S	
	H	+	-
False Positive	+	TP	FN
	-	FP	TN

→ False Negative
 → True Negative

$$TP + FN + TN + FP = n$$

Оценка качества экстрактов (пример)

N	Human	System
1	+	+
2	-	+
3	+	-
4	-	-
5	+	-

	S	
H	+	-
+	1	2
-	1	1

- Точность = $1/2$
- Полнота = $1/3$

Конференция по аннотированию текстов: DUC (Document Understanding Conference)

Формирование тестовых массивов

- Корпус текстов с аннотациями, сделанными людьми
- Корпус блоков текстов на одну тему, например, сообщения разных информагентств на тему «Визит Клинтона в Москву». Отмечены важные предложения
- Аннотирование по многим документам. Приложены примеры аннотаций 100, 200, 300, 400 слов
- Тестирование обзорных рефератов сложно, поскольку в разных текстах имеются частично похожие предложения

Автоматические ROUGE метрики

- **ROUGE** или **Recall-Oriented Understudy for Gisting Evaluation** – набор метрик и комплекс программ для оценки автоматического аннотирования и машинного перевода текстов.
- Основная идея – сравнение генерированного текста с “эталонным”, сделанным человеком.
- Существуют различные формы метрики, сравнивающие:
 1. n-граммы (ROUGE-N)
 2. минимальные общие подстроки (ROUGE-L и ROUGE-W)
 3. униграммы и биграммы (ROUGE-1 and ROUGE-2)

Автоматические ROUGE метрики

- **Общая формула:**

$$ROUGE - N(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij}))}$$

- A_i – оцениваемая обзорная аннотация i -того кластера.
- M_{ij} – ручные аннотации i -того кластера.
- $Ngram(D)$ – множество всех n -грамм из лемм соответствующего документа D .

- **Пример Rouge для двух предложений:**

1. Китай и Тайвань установили авиасообщение после 60-летнего перерыва.
2. После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем.

$$Rouge-1 = 7/12 = 0.58$$

Метод «Пирамиды» - 1

Pyramid Evaluation

- Разработан в 2005 году Колумбийским университетом.
- Эксперты выделяют из «эталонных» аннотаций «информационные единицы» - Summary Content Units (SCUs).
- Каждый SCU получает вес, равный количеству «эталонных» аннотаций, где она встречалась.
- Оценка – суммарный вес входящих SCU.
- Неоднократное вхождение SCU в автоматическую аннотацию не поощряется.

Метод «Пирамиды» - 2

- **Итоговый результат:**

[Суммарный вес найденных SCU] /

[Суммарный вес всех определённых SCU для данного топка]

- **Пример:**

SCU: Мини-субмарина попала в ловушку под водой.

1. мини-субмарина... была затоплена... на дне моря...
2. маленькая... субмарина... затоплена... на глубине 625 футов.
3. мини-субмарина попала в ловушку... ниже уровня моря.
4. маленькая... субмарина... затоплена... на дне морском...

Ручная оценка результатов

- Каждая автоматическая аннотация была прочитана несколькими экспертами NIST.
- Две оценки:
 - Содержание
 - Читабельность
- Пятибалльная система оценка – от 1 до 5.
- Результаты – заметный разрыв между автоматическими и «эталонными» аннотациями.

Сравнение методов оценки

- **ROUGE:**

- + Малое участие человека, лёгкость применения
- Отсутствие оценки читабельности, результат не всегда идеален с точки зрения человека

- **Метод «Пирамиды»:**

- + Наиболее объективная оценка содержания аннотации
- Отсутствие оценки читабельности, большое участие человека

- **Ручная оценка:**

- + Оценка «пользователем», лучшая оценка читабельности
- Огромное участие человека

Постобработка предложений

□ После отбора предложений производится улучшение связности и читаемости аннотации:

1. Замена аббревиатур

2. Привидение номеров и дат к стандартному виду

3. Замена временных ссылок:

«в конце следующего года» → «в конце 2010»

4. Замена двусмысленностей и дискурсивных форм:

«Но, это значит...» → «Это значит...»

5. Конечная сортировка предложений

Абстрагирование

- Порождение нового текста
 - В настоящее время возможно для узких предметных областей
 - Производится извлечение информации
 - Заполняются шаблоны

CBA: Concept-based Abstracting (Paice&Jones'93)

- Абстрагирование на основе шаблонов
- Извлечение важных данных из текстов: crop husbandry
 - SPECIES (the crop in the study)
 - CULTIVAR (variety studied)
 - HIGH-LEVEL-PROPERTY (specific property studied of the cultivar, e.g. yield, growth)
 - PEST (the pest that attacks the cultivar)
 - AGENT (chemical or biological agent applied)
 - LOCALITY (where the study was conducted)
 - TIME (years of the study)
 - SOIL (description of the soil)

Абстрагирование на основе шаблонов

- Шаблоны

“fertilized with *procymidane*” - шаблон

“fertilized with AGENT”

decease of SPECIES

effect of ? in SPECIES

Переменные внутри шаблона и на
концах шаблона

Порождение абстракта

- Canned-text based generation
 - this paper studies the effect of [AGENT] on the [HLP] of [SPECIES]
 - this paper studies the effect of [METHOD] on the [HLP] of [SPECIES] when it is infested by [PEST]...

Summary: *This paper studies the effect of G. pallida on the yield of potato. An experiment in 1985 and 1986 at York was undertaken.*

Несвязные аннотации

Порождение заголовков

Banko&al'00

- Порождение аннотации короче предложения
 - Text: Acclaimed Spanish soprano de los Angeles dies in Madrid after a long illness.
 - Summary: de Los Angeles died
- Порождение предложения из фрагментов разных предложений
 - Text: Spanish soprano de los Angeles dies. She was 81.
 - Summary: de Los Angeles dies at 81

Порождение заголовков

- Отбор слов
 - how many and what words to select from document
- Комбинирование слов
 - how to put words in the appropriate sequence in the headline such that it looks ok
- training: available texts + headlines

Пример

President Clinton met with his top Mideast adviser, including Secretary of State Madeleine Albright and U.S. peace envoy Dennis Ross, in preparation for a session with Israel Prime Minister Benjamin Netanyahu tomorrow. Palestinian leader Yasser Arafat is to meet with Clinton later this week. Published reports in Israel say Netanyahu will warn Clinton that Israel can't withdraw from more than nine percent of the West Bank in its next scheduled pullback, although Clinton wants 12-15 percent pullback.

- original title: *U.S. pushes for mideast peace*
- automatic title
 - *clinton*
 - *clinton wants*
 - *clinton netanyahu arafat*
 - *clinton to mideast peace*

Рамблер: Тренды (актуально)


Рамблер-Новости - Mozilla Firefox

Файл Правка Вид Журнал Закладки Инструменты Справка

louk_nat@mail.ru: Входящие Рамблер-Новости Paraphrases from Barzilay and L... В пятиэтажке в Бронницах прои... +

news.rambler.ru Яндекс


[Главная](#) [Почта](#) **Новости** [Карты](#) [Финансы](#) [Спорт](#) [Discovery](#) [ещё](#) ▼

 **Rambler**
НОВОСТИ

Главное [Политика](#) [Мир](#) [Экономика](#) [Общество](#) [Происшествия](#) [Спорт](#) [Авто](#) [Технологии](#)

Актуально: [Александр Литвиненко](#) [Братск](#) [Юлия Тимошенко](#) [Назрань](#) [Трипо](#)

сейчас [6 часов назад](#) [12 часов назад](#) [вчера](#)



[СК РФ предложил создать финансовую полицию](#)

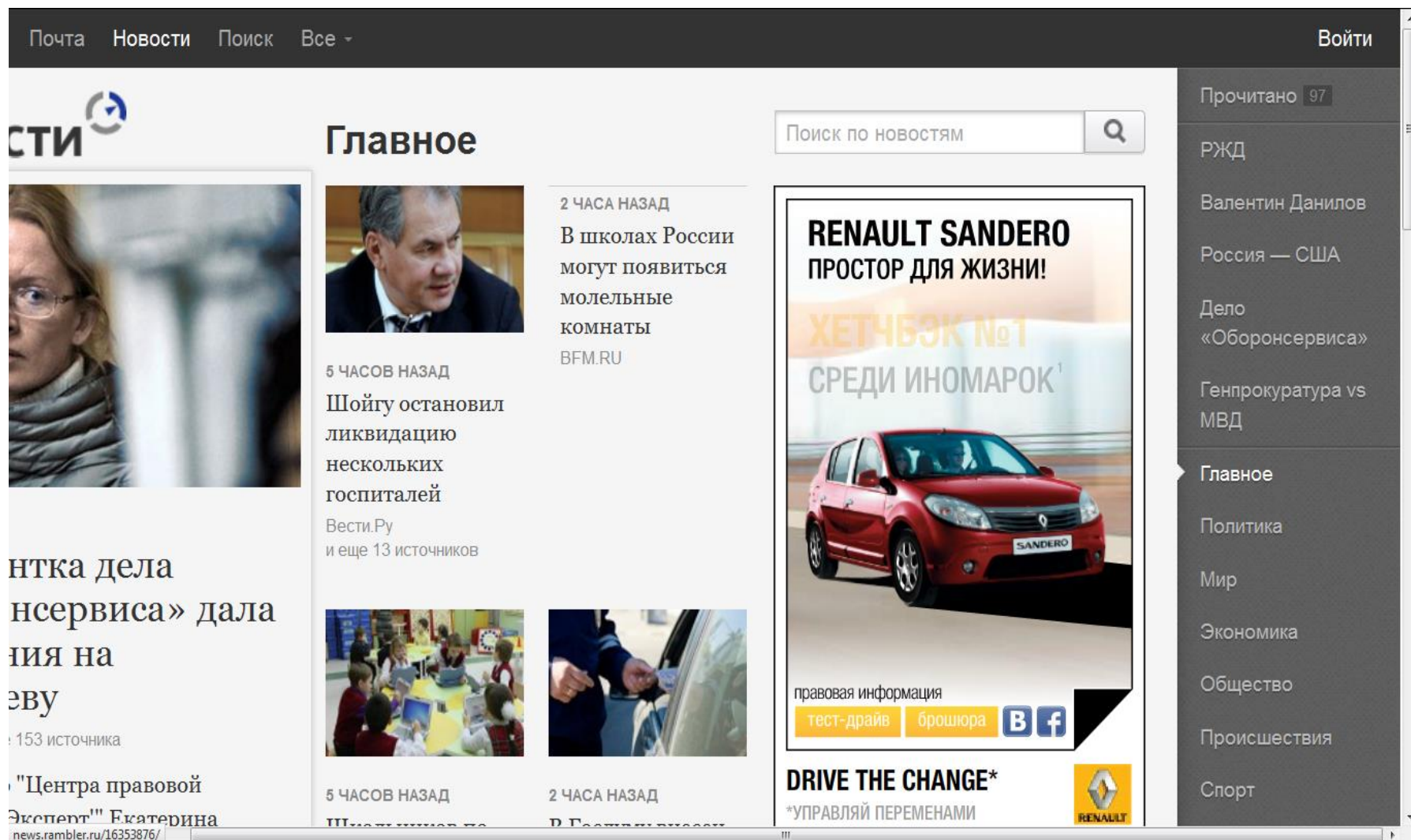
[Актер Дмитрий Дюжев на ВМ сбил школьницу](#)

[В Белом море спасатели обнаружили «робинзона»](#)

× Найти: Следующее Предыдущее Подсветить все Учсть регистр Фраза не найдена

Рамблер: тренды.

Следующая версия интерфейса



Заключение

- Виды автоматических аннотаций
- Методы порождения автоматических аннотаций
- Методы тестирования автоматических аннотаций

Применение метрики Rouge

- Эксперт включил в аннотацию предложение
 - "Русгидро", "Росгеология", "Транснефть" получат предложения перенести свои главные офисы на Дальний Восток
- Системы извлекли предложения
 - Компаниям «Русгидро», «Транснефть» и «Росгеология» предложили подумать о переезде на Дальний Восток.
 - По словам вице-премьера по Дальнему Востоку переезд может как-то затронуть "РусГидро", "Транснефть" и "Росгеологию"
 - Посчитать Rouge

$$ROUGE - N(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij}))}$$

