Автоматическое реферирование

Павел Браславский

использованы слайды:
Saggion, Medelyan, Filippova, Turpin

План

1. Автоматическое реферирование
   - Задачи, типология
   - Методы
   - Оценка
2. Веб-сниппеты
   - Задача
   - Методы
   - Производительность
   - Оценка

ЗАДАЧА

Задача

Сформировать сжатое представление (текстового) документа (связанных документов), передающее основное содержание.
+ с учетом контекста/ особенностей пользователя.
+ связное и удобочитаемое.

FEATURE | WSJ. MAGAZINE 2013 TECHNOLOGY INNOVATOR

How Teen Nick D'Aloisio Has Changed the Way We Read

When a Hong Kong billionaire emailed a London tech startup to inquire about investing, he didn't realize its e
consisted of a single kid working in his bedroom. Meet the 18-year-old WSJ. Magazine Technology Innovator
became an overnight millionaire by inventing an app that may revolutionize how we read on the go

CME Group
energy com
manage the
and plan fo
LEARN MO
How the wo
CME

Не только новости

- Научные статьи
- Художественные произведения
- Отзывы о товарах
- Веб-страницы (сниппеты – см. ниже)
- …

Актуальность

- Большие объемы информации
- Быстрое принятие решений
- Мобильные устройства

Не только текст

- Кино
- Музыка
- Видео
- Речь

ПРИМЕРЫ

ELSEVIER

Information Processing and Management 40 (2004) 919–938

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Centroid-based summarization of multiple documents

Dragomir R. Radev [a,*], Hongyan Jing [b], Małgorzata Styś [b], Daniel Tam [a]

[a] *University of Michigan, Ann Arbor, MI 48109, USA*

[b] *IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA*

Received 5 January 2003; accepted 24 October 2003

Available online 13 December 2003

Abstract

We present a multi-document summarizer, MEAD, which generates summaries using cluster centroids produced by a topic detection and tracking system. We describe two new techniques, a centroid-based summarizer, and an evaluation scheme based on sentence utility and subsumption. We have applied this evaluation to both single and multiple document summaries. Finally, we describe two user studies that test our models of multi-document summarization.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Multi-document summarization; Centroid-based summarization; Cluster-based relative utility; Cross-sentence informational subsumption

Яндекс Новости

Главные новости | Мои новости | Политика | Общество | Экономика | В мире | Спорт | Происше

Миссия для поиска жизни на Марсе успешно стартовала с Байконура

Аргументы и Факты 15:40

Миссия для поиска жизни на Марсе успешно стартовала с Байконура

С космодрома «Байконур» успешно стартовала миссия для поиска жизни на Марсе «ЭкзоМарс-2016» (ExoMars-2016), передает РИА Новости. Сейчас разгонный блок «Бриз-М» отделился от ракеты-носителя.

Lenta.ru 14:34

Стартовала ракета «Протон-М» с модулями проекта ExoMars

Ракета «Протон-М» успешно стартовала с модулями проекта ExoMars. Трансляция запуска проходит в прямом эфире на YouTube-канале Роскосмоса. Первая в истории совместная российско-европейская миссия ExoMars-2016 предполагает отправку к Марсу орбитального и демонстрационного спускаемого модулей.

ТАСС 12:56

Французский ученый: проект «ЭкзоМарс» - важный вклад в космическую науку Европы и России

статья

Цель - поиск метана в атмосфере Марса Ученый отметил, что важной частью программы "ЭкзоМарс" является поиск в его атмосфере следов метана марсианским научным орбитальным аппаратом Trace Gas Orbiter.

Свернуть все сообщения

Екатеринбург Он-лайн 16:46

С Байконура стартовала миссия для поиска жизни на Марсе

Ракета с межпланетной станцией успешно стартовала с Байконура. Сегодня днём ракета "Протон-М" успешно стартовала с космодрома Байконур и вывела на околоземную орбиту космические аппараты российско-европейской миссии ExoMars-2016.

Московский Комсомолец 15:15

Ракета-носитель «Протон-М» успешно стартовала к Марсу с

Яндекс.Новости

Лучший заголовок

Экономика // Франция

Франция не рассматривает вариант продажи Societe Generale

Также официальный представитель администрации президента Франции отказался комментировать информацию о возможной заинтересованности другого французского банка *BNP Paribas* в покупке *Societe Generale*, отметив лишь, что это касается акционеров обоих банков. Газета.Ru 18:56

Французская полиция провела обыск в квартире брата трейдера *Societe Generale* Жерома Кервьеля, последний подозревается в мошенничестве, в результате которого банк потерял 5 миллиардов евро. Российская газета 10:18

Французская полиция в субботу взяла под стражу трейдера *Жерома Кервьеля*, из-за которого банковская группа *Societe Generale* понесла потери. NEWSru Украина 30.01 13:57

Еще на эту тему

Министр экономики Франции намеренно скрыла скандал в банке 29.01 20:22

Французы покупают "Росбанк" 15:21

"Мы приобретаем второй по числу отделений банк России", - заявил глава банка Societe Generale Даниэль Бутон , выступая в прямом эфире французского телевидения.

Даниэль Бутон, Societe Generale, глава все цитаты сюжета (135)

В сюжете: Жером Кервьель, Даниэль Бутон, Николя Саркози, Жан-Клод Марэн, Кристиан Шарьер-Бурназель • Societe Generale

Краткие выдержки

Цитата

Персоны

Фото (всего 47)

Правда.ru

Утро.ru

Видео (всего 5)

Вести.Ru

РИА Новости

Аудио (всего 3)

Радио России

Русская Служба Новостей

Франция не рассматривает вариант продажи Societe Generale Газета.Ru 18:56

Продажа банка Societe Generale иностранному или французскому банку-конкуренту "не ... Прайм-ТАСС 18:56

Вариант продажи Societe Generale не рассматривается - Мартинон РИА Новости 18:37

BNP Paribas рассматривает возможность поглощения Societe Generale K2Kapital.com 18:03

SG не собирается сворачивать бизнес в РФ, несмотря на финансовые потери Интерфакс 17:01

Французский банк может купить Societe Generale Правда.ru 16:57

Societe Generale: скандал во Франции не повлиял на бизнес России NEWSru.com 15:56

UniCredit не заинтересован в покупке Societe Generale Finam.Ru 15:41

Французский банк BNP Paribas может купить Societe Generale Коммерсантъ-Online 15:09

Societe Generale в сфере внимания BNP Paribas Advisers 15:08

Статьи на эту тему (всего 25)

Подборка

Гримасы капитализации Коммерсантъ 28.01 01:49

Societe Generale попал на 50 млрд евро КоммерсантЪ (Украина) 28.01 09:00

Большие банки громко лопаются Советская Белоруссия 29.01 12:30

Французский генерал списывает поражение на рядового? Московский комсомолец 29.01 15:45

Откровения Кервьеля Время новостей 30.01 09:45

Главные новости

Число жертв взрыва на «Северстали» может возрасти

Генпрокуратура обжаловала перенос дела «Трех китов» в Петербург

Число жертв взрыва здания Делового центра в Стамбуле возросло до 20

У бывшего вице-президента ЮКОСа Василия Алексаняна обнаружена лимфома

ВЦИОМ: Медведев выглядит очень позитивно в глазах населения

Экономика

Зубков станет преемником преемника

Стабфонд прекратил свое существование

Северсталь продает угольные активы в Кемеровской области

Регион: Екатеринбург

(495)

Лабиринт

Корзина

Жанр: Книги / Художественная литература / Русская проза / Классика

Война и мир: в 4 томах. Том 1-4

ID товара: 186469
ISBN: 978-5-699-31874-2

Автор: Толстой Лев
Издательство: Эксмо
Серия: Большая книга
Жанр: Классика
Год выпуска: 2008

Страниц: 1296 (Офсет)
Тип обложки: 7Бц - твердая целлофанированная (или ла
Иллюстрации: Без иллюстра

Масса: 1146 г
Размеры: 218x147x56 мм

Наличие: **На складе (огра количество)**

339 руб. Купить

Метки: Добавить метку

Регион: Екатеринбург

(495) 733-91-74

279041902

Лабиринт

Корзина

Доставка

Жанр: Книги / Учебная, методическая литература и словари / Книги для школы / Литература / Краткое изложение произведений

"Война и мир" Л. Н. Толстого. Краткое содержание. Особенности романа. Сочинения

ID товара: 72004
ISBN: 5-8053-0442-2

Автор: Шаповалова Оксана
Издательство: Лицей
Серия: По страницам литературной классики
Жанр: Краткое изложение произведений
Год выпуска: 2005

Страниц: 304 (Газетная)
Тип обложки: обл - мягкий переплет (крепление скрепкой или клеем)
Иллюстрации: Черно-белые

Масса: 276 г
Размеры: 200x140x15 мм

Наличие: **На складе**

Содержание

92 руб. Купить

Метки: Добавить метку

анные
прикладные задачи...

Поиск Почта Новости Маркет Карты Словари Блоги Картинки еще ▼

Яндекс

Нашлось
9 млн страниц

школа анализа данных

☐ в найденном ☐ в Екатеринбурге расшир

1. **Школа анализа данных** и информатики
   О **Школе**. Как поступить? Программа. Преподаватели. Расписание. Вопросы и ответы. Сотрудники. **Школа анализа данных** и информатики. **Школа анализа данных** и информатики придумана, организована и профинансирована компанией «Яндекс» в 2007 году.
   company.yandex.ru/school
   Сохраненная копия · Еще с сайта 2865 · Рубрика: Поисковые системы · Это спам

2. Новости о грантах. Объявления и информация.
   Школа анализа данных придумана, организована и профинансирована компанией «Яндекс» в 2007 году. Цель **школы** — подготовка специалистов в области **анализа данных** и извлечения информации из интернета, проведение фундаментальных научных разработок в этой области и участие в прикладных проектах Яндекса.
   www.rsci.ru/MoreInfo.html?MessageID=6240 · 18 КБ
   Сохраненная копия · Еще с сайта 70 · Рубрика: Фонды, стипендии, гранты · Это спам

3. :: ИКТ Портал :: Новости
   Школа анализа данных компании "Яндекс" / 02.06.2008. **Школа анализа данных** придумана, организована и профинансирована компанией "Яндекс" в 2007 году.
   www.ict.edu.ru/news/study/2046/ · 10 КБ
   Сохраненная копия · Еще с сайта 461 · Рубрика: Информатика, информационные системы · Это спам

4. "Яндекс" набрал первый курс в "**Школу анализа данных**" | Вебпланета
   Все вместе | Новости | Пресс-релизы | Интервью | Ноу-хау | Обзоры | Репортажи | Мнения | Авторы | Читальный зал | **Школа** | Вопросы и ответы | О нас | Контакты. Вход Регистрация. Пользовательский логин.
   ...**Школе анализа данных**" Яндекса. "Яндекс" представляла привычная тройка: Аркадий Волож, генеральный директор, Елена Колмановская, главный...
   www.webplanet.ru/news/life/2007/09/21/yandex_school.html · 36 КБ
   Сохраненная копия · Еще с сайта 2589 · Рубрика: Интернет · Это спам

5. От переработки **данных** - к **анализу** | Статьи | **Школа** инвестора | База знаний
   Главная > База знаний > **Школа** инвестора > Статьи > От переработки **данных** - к **анализу**.
   know.onlinebroker.ru/articles/rebild2analize.asp · 54 КБ
   Сохраненная копия · Еще с сайта 5070 · Рубрика: Брокеры и управление активами · Это спам

6. Яндекс приглашает студентов в **Школу анализа данных**. Новости от Searchengines.ru

Типы рефератов

- Abstract ←→ Extract
- Индикативные ←→ информативные
- Один ←→ много документов
- Общие ←→ с учетом контекста/пользователя
- Критический

МЕТОДЫ

Методы

- Одна из самых ранних задач ОЕЯ
- Как и в других задачах, подходы knowledge/rule-based vs. статистика

Abstract/Extract

- Текст → «семантическое представление» (сценарии/фреймы) → генерация реферата

(abstract generation)

- Выбор значимых фрагментов, обычно предложений

(sentence extraction)

Экстракты

- + простота
- - связность
  - дисбаланс (ср. разные точки зрения в статье)

Ранние работы

- Keyword distribution (Luhn’58)
- Position Method (Edmundson’69)
- Title Method (Edmundson’69)
- Cue Method/Indicative Phrases (Edmundson’69; Paice’81)
- DeJong (1979, 1982)

[Saggion]

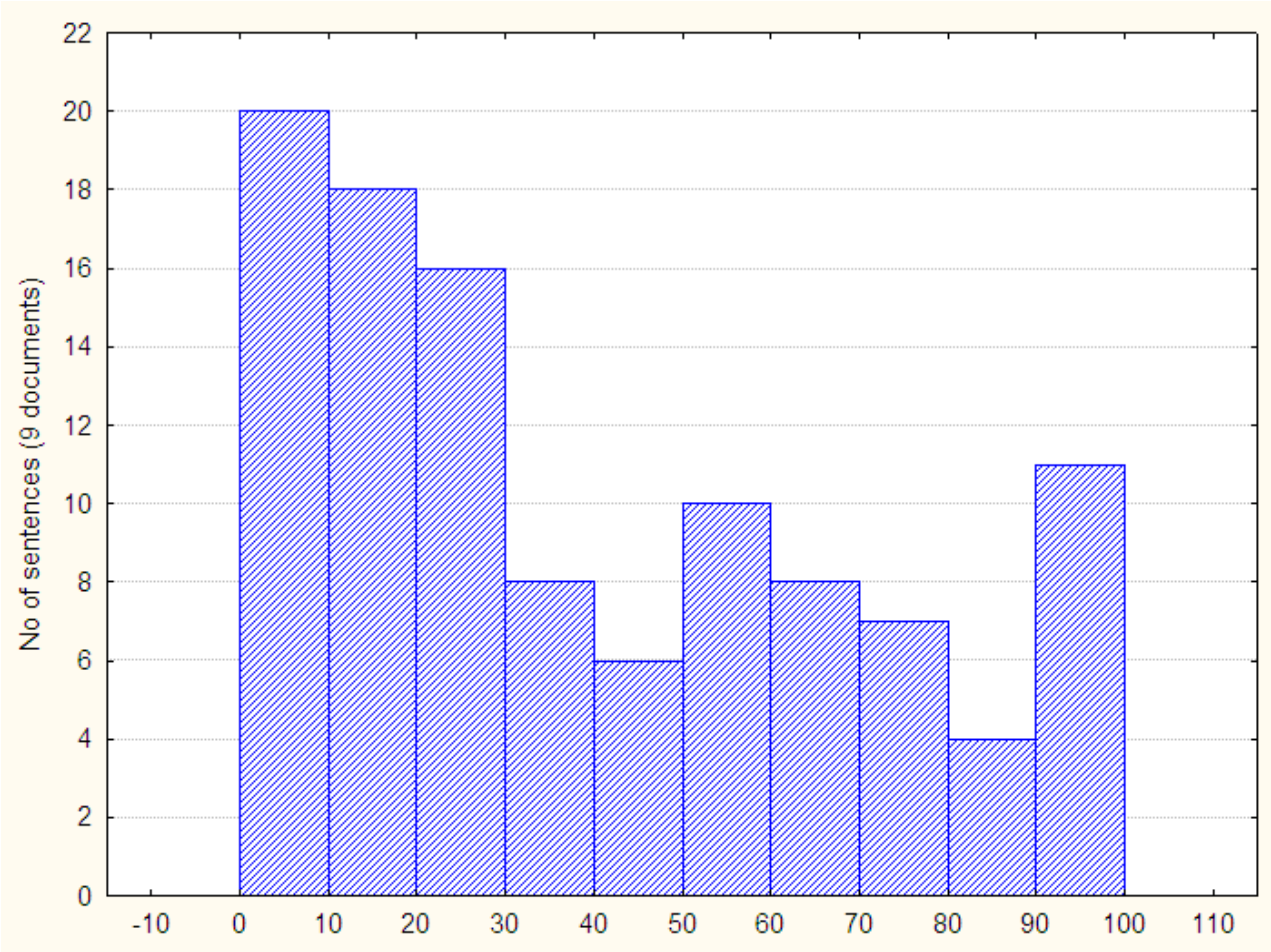
Слова

- Вес предложения пропорционален весу входящих в него слов.
- Вес слова вычисляется на основе частоты
- Близко расположенные важные слова усиливают друг друга (*keyword burst*)

Предложения

- Начало и конец документа/абзаца/раздела обычно содержат «хорошие» предложения (если текст «хороший»)
- Похоже на заголовок – хорошо
- Нормализация по длине предложения: длинные предложения набирают вес просто за счет длины, короткие предложения часто неинформативны
- Сигнальные фразы и слова: *можно сделать вывод…*, *необходимо подчеркнуть…* (ограничение по жанру, сложно поддерживать словарь)
- Вопросительные предложения скорее всего не подходят
- Предложения с анафорической отсылкой не подходят (или включаются в реферат вместе с предыдущим предложением)

Положение предложений



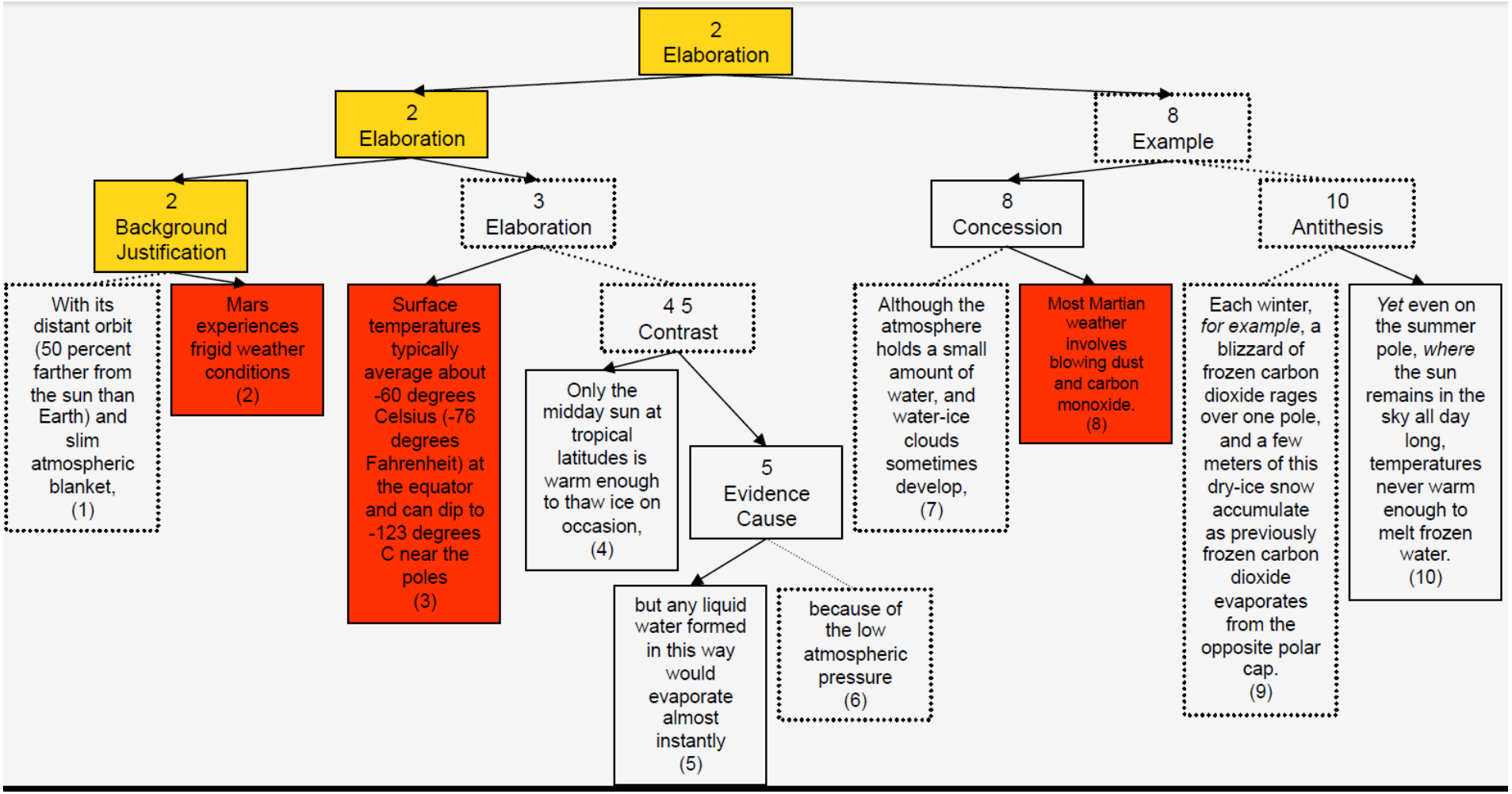
Положение выделенных вручную «информативных» предложений (9 новостных текстов, каждый обрабатывался 3 асессорами, длины текстов нормированы)

Комбинация

- [Edmundson’69]

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

AS via RST (Marcu,1997-1999)



Relation names (in M&T 1988)

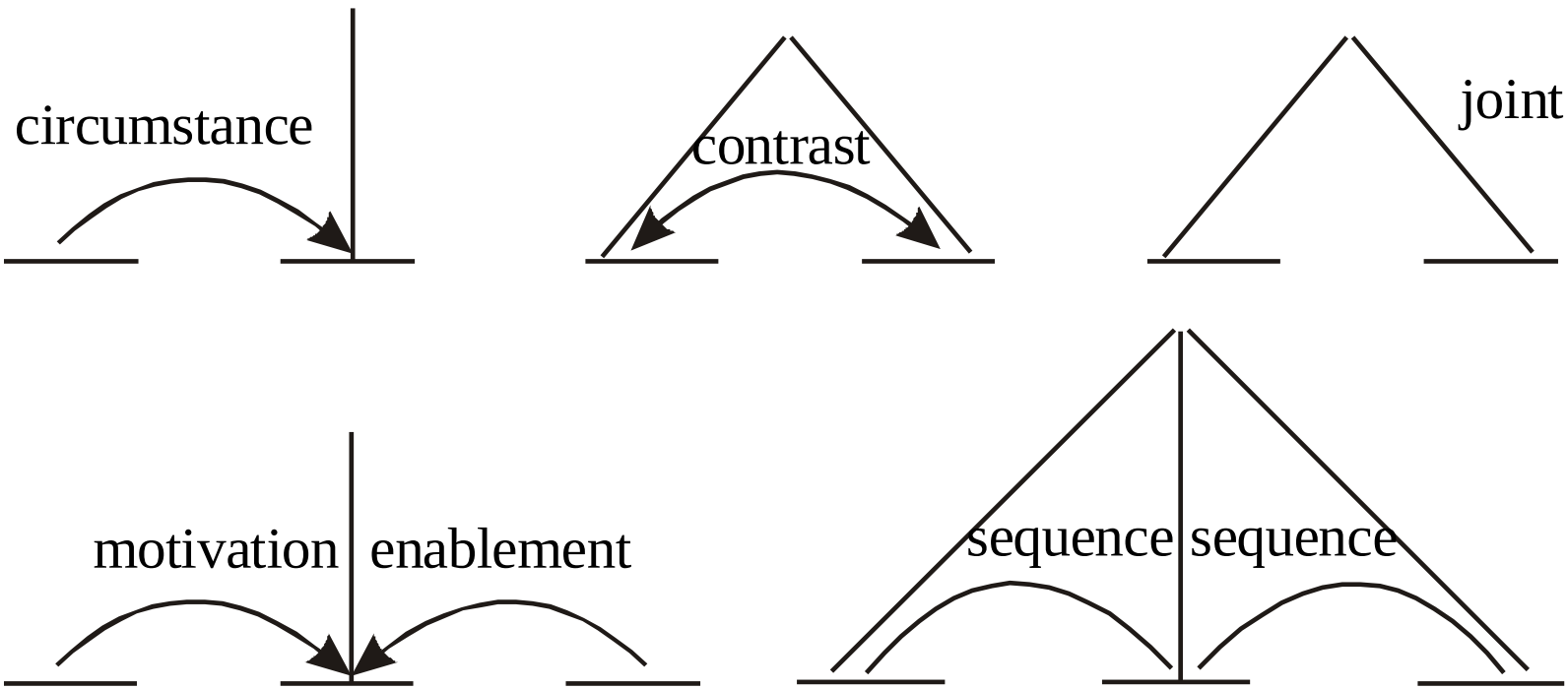
Circumstance	**Antithesis and Concession**
Solutionhood	**Antithesis**
Elaboration	**Concession**
Background	**Condition and Otherwise**
Enablement and Motivation	**Condition**
Enablement	**Otherwise**
Motivation	**Interpretation and Evaluation**
Evidence and Justify	**Interpretation**
Evidence	**Evaluation**
Justify	**Restatement and Summary**
Relations of Cause	**Restatement**
Volitional Cause	**Summary**
Non-Volitional Cause	**Other Relations**
Volitional Result	**Sequence**
Non-Volitional Result	**Contrast**
Purpose	

Other classifications are possible, and longer and shorter lists have been proposed

Taboada & Stede, 2009

Schemas

- They specify how spans of text can co-occur, determining possible RST text structures



Taboada & Stede, 2009

ML4AS (Kupiec’95)

- Корпус статей с рефератами для обучения
- Признаки:
  - длина, положение предложения
  - присутствие «тематических» слов, слов из заглавных букв, «ключевых» фраз
- Сопоставление предложений из реферата и основного текста
- Байесовский классификатор

A Trainable Document Summarizer

Julian Kupiec, **Jan Pedersen** and **Francine Chen**
Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
{kupiec,pedersen,fchen}@parc.xerox.com

Abstract

- To summarize is to reduce in complexity, and hence in length, while retaining some of the essential qualities of the original.
- This paper focusses on document extracts, a particular kind of computed document summary.
- Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document, which suggests that even shorter extracts may be useful indicative summaries.
- The trends in our results are in agreement with those of Edmundson who used a subjectively weighted combination of features as opposed to training the feature weights using a corpus.
- We have developed a trainable summarization program that is grounded in a sound statistical framework.

Keywords: summary sentence, original documents, summary pairs, training corpus, document extracts

author-supplied indicative abst
it is hoped that other, more easi
serve.

Numerous researchers have
marization (see [10] for an ove
ing a coherent narrative summ
sidered too problematic since i
ing, abstraction, and language
edge intensive methods have ha
[11, 5, 3, 13, 18]. For example,
sage understanding system ca
ment summary. A simpler, mo
tral difficulties of natural langu
to be *summary by extraction* [7
of the document that is indicat
ing sentences and presenting th
of summaries are not guarante
may be useful for rapid releva

Document extracts consisti
be as informative as the full te
that even shorter extracts may b
ever, other studies [12, 2] sug
far from unique. Numerous he

[Браславский, Густелев, 2007]

- Машинное обучение (LibSVM)
- «Дешево» получить данные
- Использовать разнообразный набор признаков

Россия и Европа полетели на Марс

На Марс отправился космический аппарат «Экзомарс» — совместный проект ЕЗА и «Роскосмоса», который исследует атмосферу Красной планеты. О том, как миссия поможет ответить на вопрос о существовании жизни на Марсе и что еще будет входить в ее задачи, рассказывает отдел науки «Газеты.Ru».

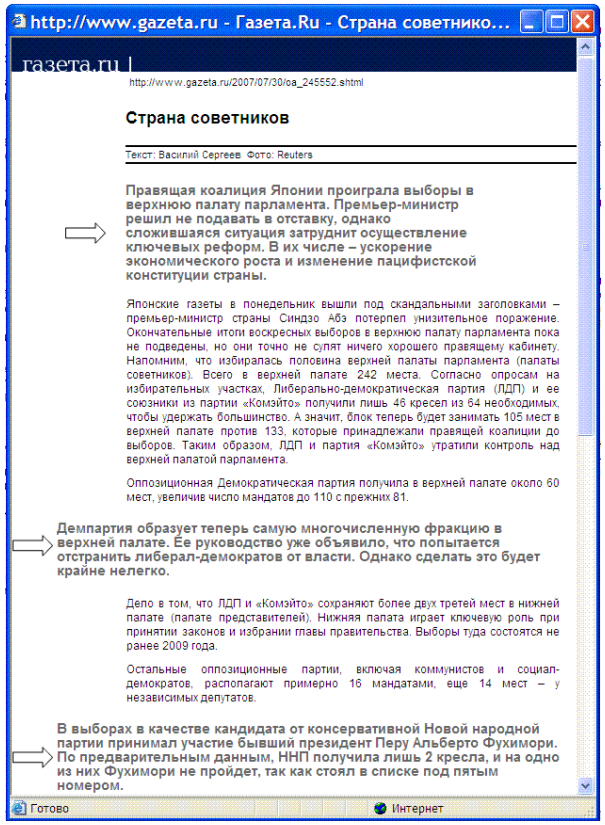
Подобно многим автоматическим станциям, «Экзомарс-2016» состоит из двух блоков, орбитального и посадочного.

После отделения модуля «Скиапарелли» орбитальный аппарат выйдет на эллиптическую орбиту. Он пролетит через верхние слои атмосферы Марса, а затем выровняет орбиту до круговой с высотой около 400 км.

на Марсе тоже протекают биологические процессы. Однако газ может иметь и химическое происхождение: он образуется в ходе различных геологических процессов.

Второй этап миссии намечен на 2018 год: в ходе него на Марс будет отправлен космический аппарат «Экзомарс-2018». Он будет состоять из трех блоков: перелетный модуль, аппарат с системой отделения и спускаемый аппарат, который разрабатывает НПО имени Лавочкина.

Данные



http://www.gazeta.ru - Газета.Ru - Страна советнико...

газета.ru |

http://www.gazeta.ru/2007/07/30/oa_245552.shtml

Страна советников

Текст: Василий Сергеев Фото: Reuters

Правящая коалиция Японии проиграла выборы в верхнюю палату парламента. Премьер-министр решил не подавать в отставку, однако сложившаяся ситуация затруднит осуществление ключевых реформ. В их числе – ускорение экономического роста и изменение пацифистской конституции страны.

Японские газеты в понедельник вышли под скандальными заголовками – премьер-министр страны Синдзо Абэ потерпел унизительное поражение. Окончательные итоги воскресных выборов в верхнюю палату парламента пока не подведены, но они точно не сулят ничего хорошего правящему кабинету. Напомним, что избиралась половина верхней палаты парламента (палаты советников). Всего в верхней палате 242 места. Согласно опросам на избирательных участках, Либерально-демократическая партия (ЛДП) и ее союзники из партии «Комэйто» получили лишь 46 кресел из 64 необходимых, чтобы удержать большинство. А значит, блок теперь будет занимать 105 мест в верхней палате против 133, которые принадлежали правящей коалиции до выборов. Таким образом, ЛДП и партия «Комэйто» утратили контроль над верхней палатой парламента.

Оппозиционная Демократическая партия получила в верхней палате около 60 мест, увеличив число мандатов до 110 с прежних 81.

Демпартия образует теперь самую многочисленную фракцию в верхней палате. Ее руководство уже объявило, что попытается отстранить либерал-демократов от власти. Однако сделать это будет крайне нелегко.

Дело в том, что ЛДП и «Комэйто» сохраняют более двух третей мест в нижней палате (палате представителей). Нижняя палата играет ключевую роль при принятии законов и избрании главы правительства. Выборы туда состоятся не ранее 2009 года.

Остальные оппозиционные партии, включая коммунистов и социал-демократов, располагают примерно 16 мандатами, еще 14 мест – у независимых депутатов.

В выборах в качестве кандидата от консервативной Новой народной партии принимал участие бывший президент Перу Альберто Фухимори. По предварительным данным, ННП получила лишь 2 кресла, и на одно из них Фухимори не пройдет, так как стоял в списке под пятым номером.

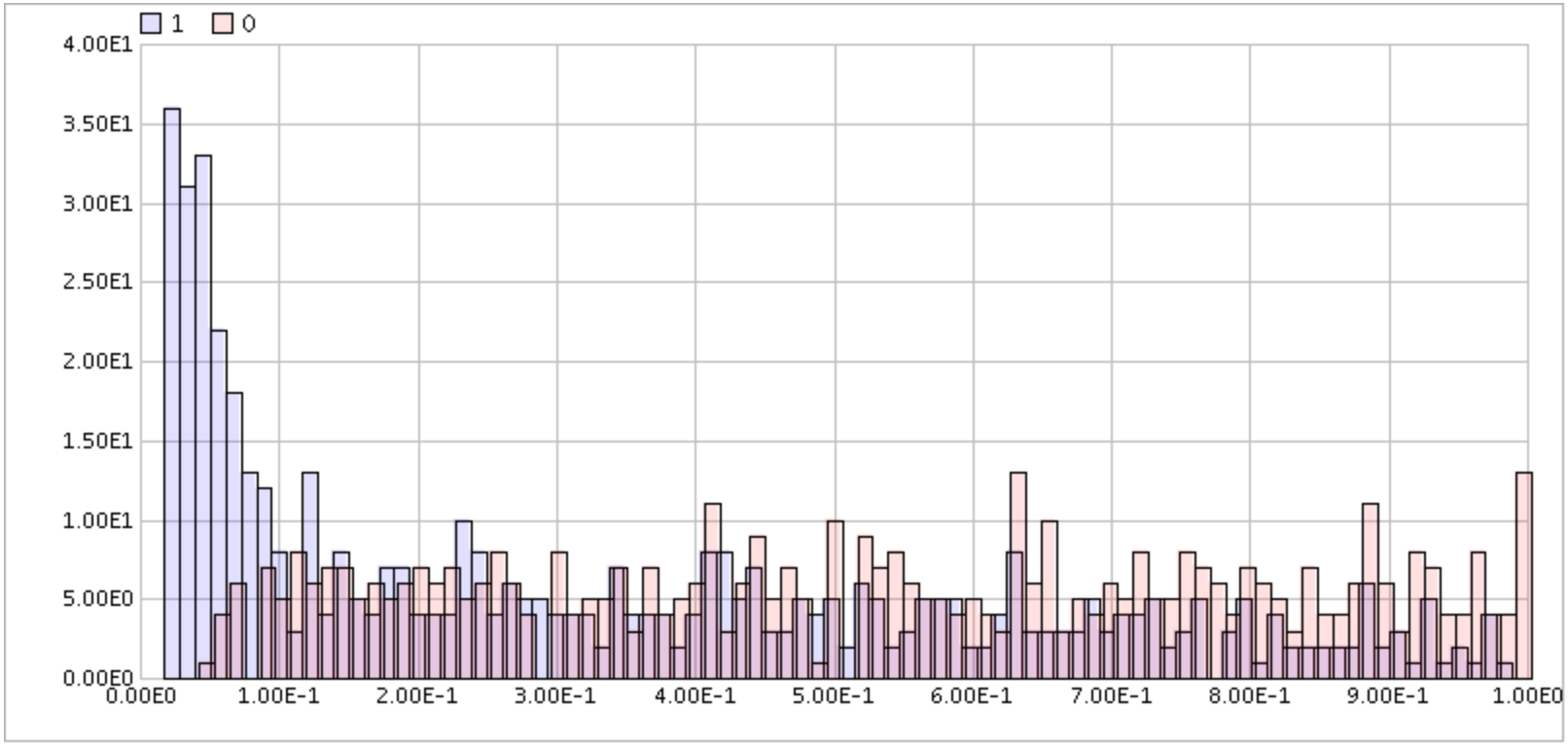
Готово Интернет

		Доку-менты	Пред-ложения	Квазиреферат
ОВ	Бизнес	294	9534	1870 / 19,6%
	Политика	570	21153	4272 / 20,1%
	Всего	864	30687	6142 / 20,0%
ТВ	Бизнес	89	2843	540 / 18,8%
	Политика	230	8753	1692 / 19,3%
	Всего	319	11596	2232 / 19,2%
Всего		1183	42283	8374 / 19,8%

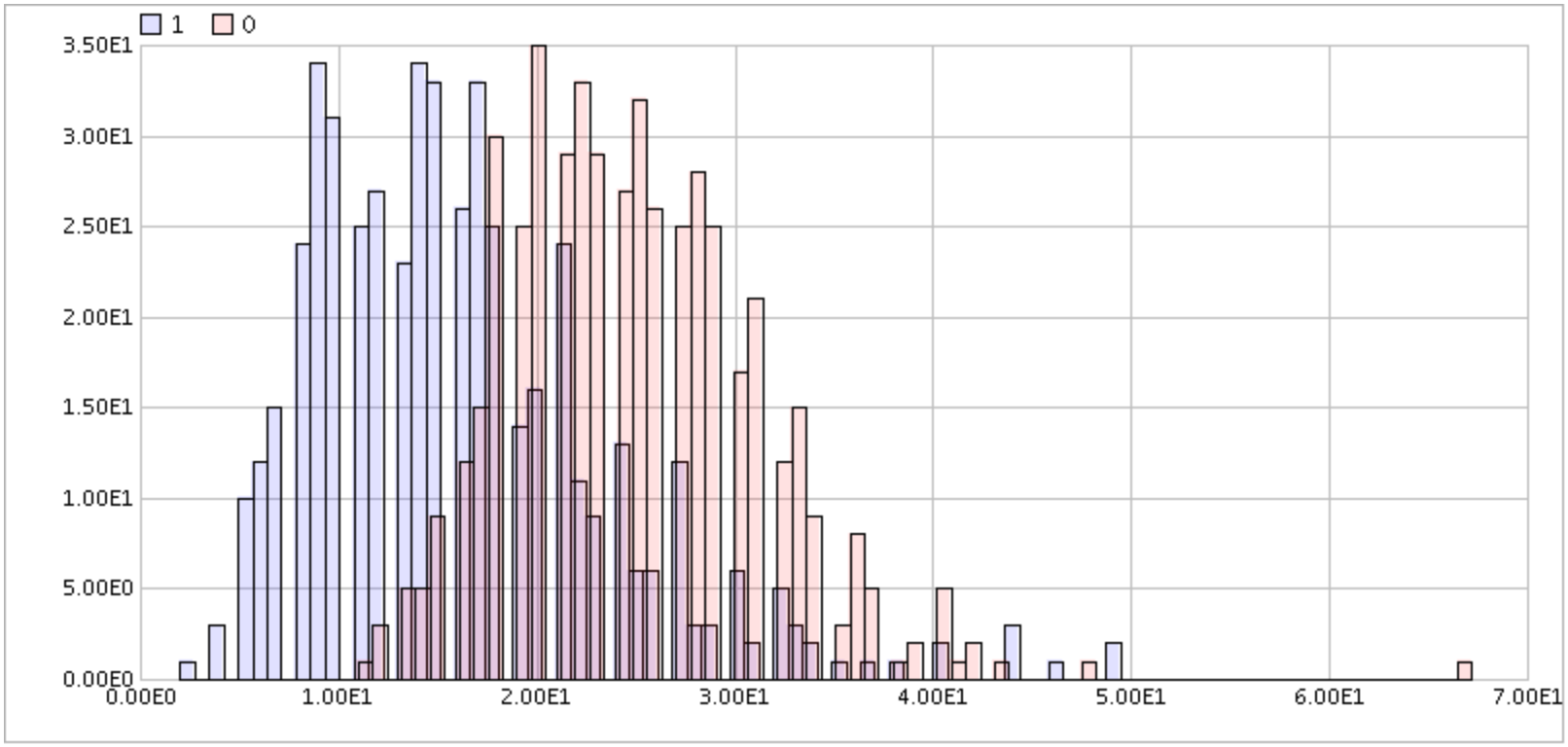
Признаки

- Тематические признаки
  - топ-10 частотных слов документа
  - сумма весов слов (TF*IDF)
  - имя + фамилия.
  - слова заголовка
- Структурные признаки предложения
  - вопросительное предложение
  - прямая речь
  - длина предложения, отнесенная к самому длинному предложению в тексте
  - подчинительный союз из списка
  - синтаксическая сложность предложения
- Структурные признаки текста
  - топ-10 частотных слов документа для предыдущего и последующего
  - сумма весов слов (TF*IDF), для предыдущего и последующего
  - простой сигнал анафоры
  - положение предложения в тексте
- Формальные признаки
  - длина предложения в знаках
  - количество запятых в предложении
  - количество точек в предложении
  - количество чисел в предложении

Признаки: пример



Признаки: пример (2)



Методика оценки

- доля совпадающих предложений на реферате той же длины
  - по тестовой выборке
  - «независимая» выборка + ручной выбор предложений
    - 30 документов, 2 эксперта, 231 предложение
    - согласие: 36% (!)

Результаты

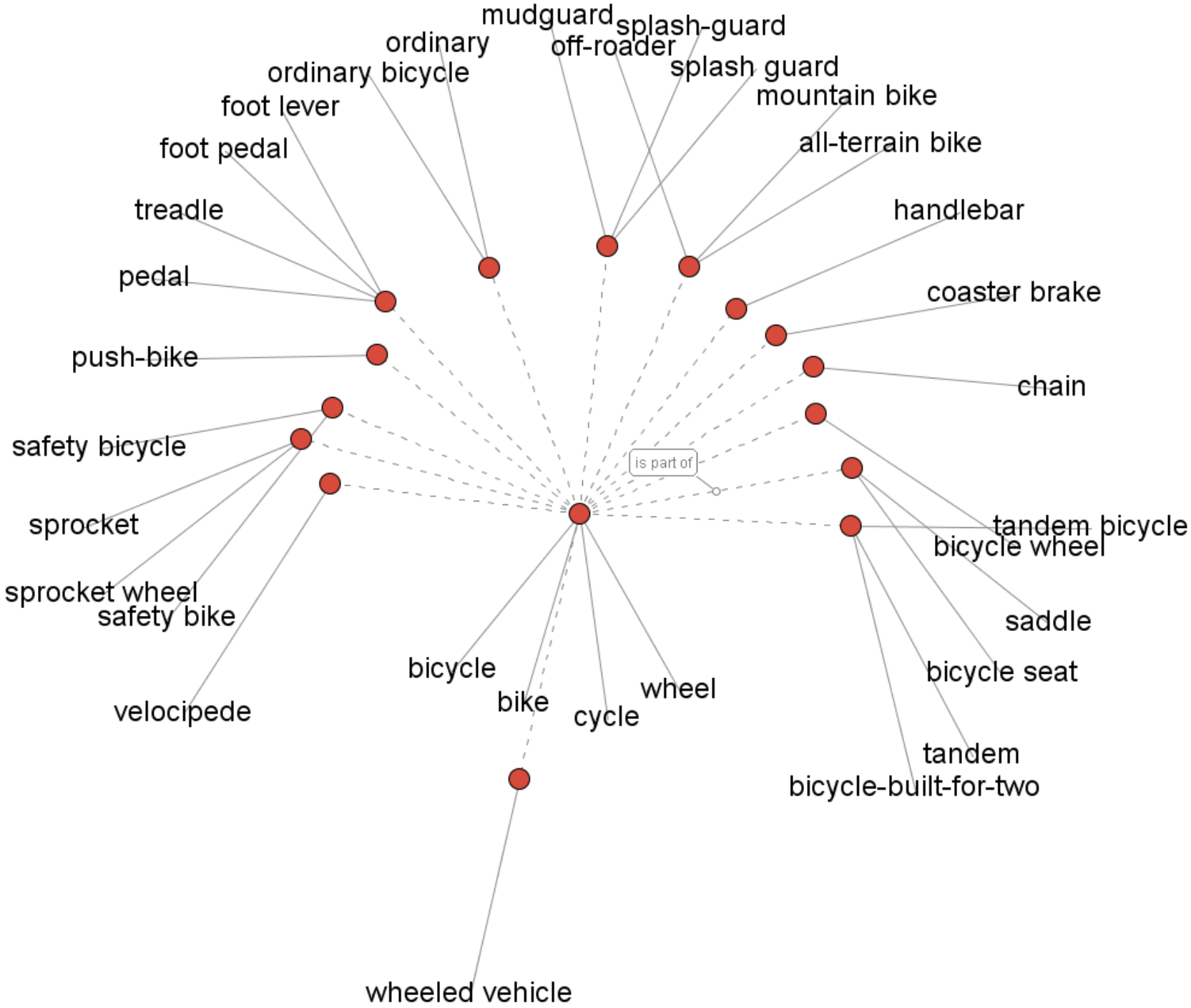
	наша система	baseline
Политика	52,83%	52,74%
Бизнес	50,28%	51,57%
Вся ТВ	52,12%	51,97%
Эксперт 1	31,16%	35,90%
Эксперт 2	51,94%	64,53%
Согласие	54,21%	74,70%
Разнообразие	36,41%	64,91%

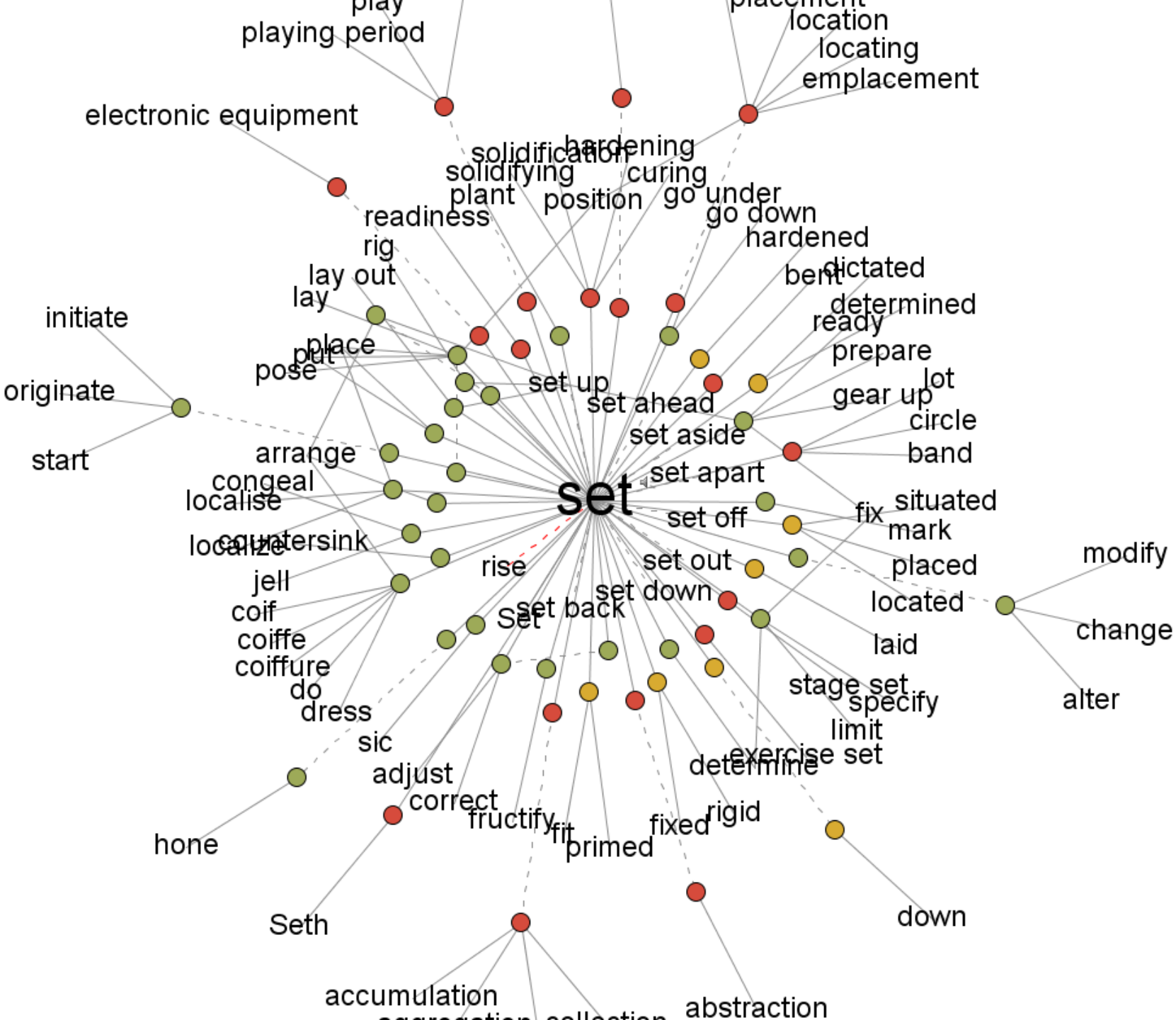
Lexical Chains

- Lexical chains are **sequences of related words:**

  {cow, sheep, wool, scarf, hat}
- Used to identify cohesive structure in a text, its discourse flow and topic areas
- **Computable with any thesaurus** containing semantic links between words
- In **NLP**:
  - used for text segmentation, text summarization, question answering, word sense disambiguation, keyphrase indexing

[Medelyan]


mudguard
splash-guard
ordinary
off-roader
ordinary bicycle
splash guard
mountain bike
foot lever
all-terrain bike
foot pedal
handlebar
treadle
pedal
coaster brake
push-bike
chain
safety bicycle
is part of
sprocket
tandem bicycle
bicycle wheel
sprocket wheel
safety bike
saddle
bicycle
bicycle seat
wheel
bike
cycle
velocipede
tandem
bicycle-built-for-two
wheeled vehicle


set
play
playing period
placement
location
locating
emplacement
electronic equipment
solidification
hardening
solidifying
curing
plant
position
go under
go down
readiness
rig
hardened
lay out
bent
dictated
lay
determined
ready
initiate
place
prepare
put
pose
set up
lot
originate
set ahead
gear up
circle
set aside
start
arrange
band
congeal
set apart
localise
set off
fix
situated
mark
countersink
localize
rise
set out
placed
modify
jell
set down
located
coif
set back
Set
change
coiffe
laid
coiffure
stage set
do
specify
alter
dress
limit
sic
exercise set
determine
adjust
correct
rigid
fructify
fit
fixed
hone
primed
Seth
down
accumulation
abstraction

How a Lexical Chain is Created

1.1.4 Impact of Natural Disasters in the Highlands of Central Vietnam
The most damaging hazard experienced in the highlands is flashflood, as it occurs with little warning. People, property and livestock may be washed away. Crops planted on the hillsides are better protected than staple crops in river valleys, such as cassava, on which poorer farmers rely between paddy harvests. High dependence on subsistence farming renders highland populations vulnerable to hunger during the flood season. Floods from swollen rivers can cut off villages for days or weeks, which could result in food shortages.
Floods with strong currents cause permanent damage to fields, washing away the topsoil. Floodwaters also deposit rock and gravel onto fields. Heavy rains trigger landslides that cut off roads and communication networks.

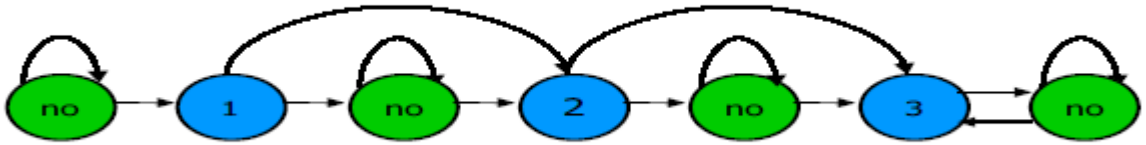
[Medelyan]

Barzilay & Elhadad (1997)

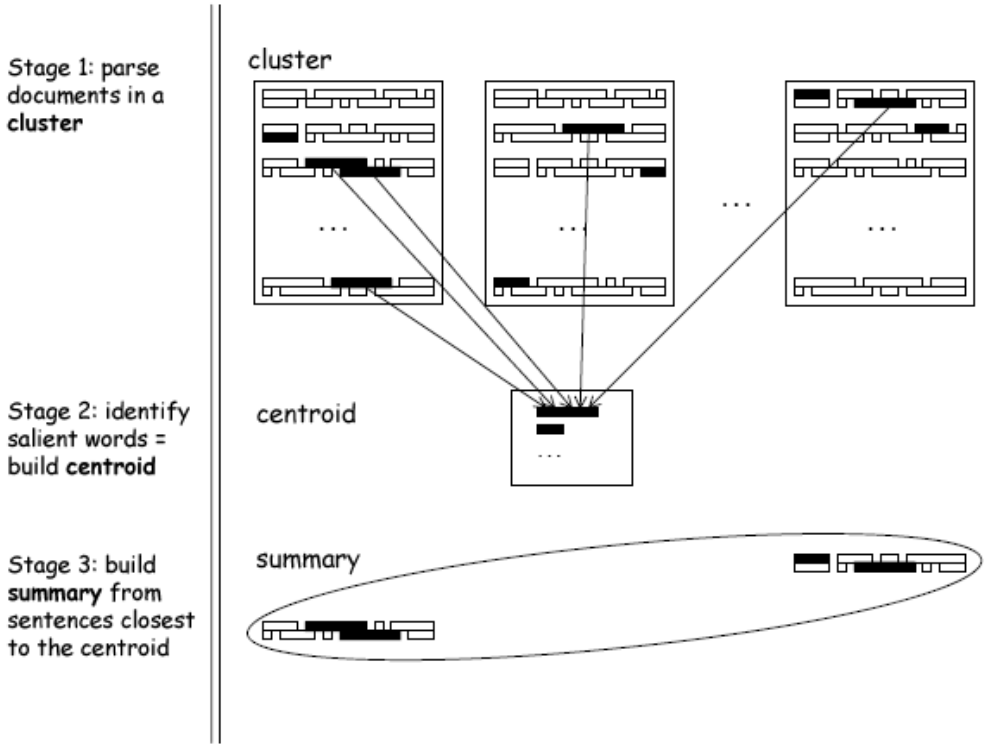
1. Рассматриваем все существительные + именные фразы (e.g. *sea level*) в качестве кандидатов
2. три типа связи: повторы слова < синсеты слов непосредственно связаны < есть путь длины 1+
3. слова должны встречаться в ограниченном окне
4. построение всех возможных цепочек, снятие неоднозначности (выбирается наиболее связный граф, разные веса для разных типов связей)
5. Взвешивание цепочек: встречаемость звеньев в тексте, встречаемость – количество уникальных звеньев → «сильные цепочки»
6. Извлечение предложений:
   1. первое появление звена каждой цепочки
   2. первое появление самого частотного звена *
   3. предложения из «насыщенного» сегмента

Реферирование на основе HMM

- [Conroy & O'Leary, 2001]
- 2s+1 состояний
- признаки:
  - Позиция (документ/абзац)
  - # терминов, статистика по коллекции



Multi-document summarization



MEAD

вес предложения:

+ слова из центроида

+ положение

+ похожесть на первое предложение

-похожесть на другие

(Radev et. al, 2004)

Графовое представление

- Близость предложений на основе коэффициента Жаккара + порог → граф предложений
- Центральность – например, на основе PageRank
- Упорядочение предложений на основе значения центральности

Борьба с избыточностью

- Maximal marginal relevance, MMR (Carbonell&Goldstein, 1998)
- Cross-sentence informational subsumption, CSIS (Radev et al., 2004)
- кластеризация предложений (Nomoto&Matsumoto, 2003)

Но: избыточность может способствовать связности и удобочитаемости

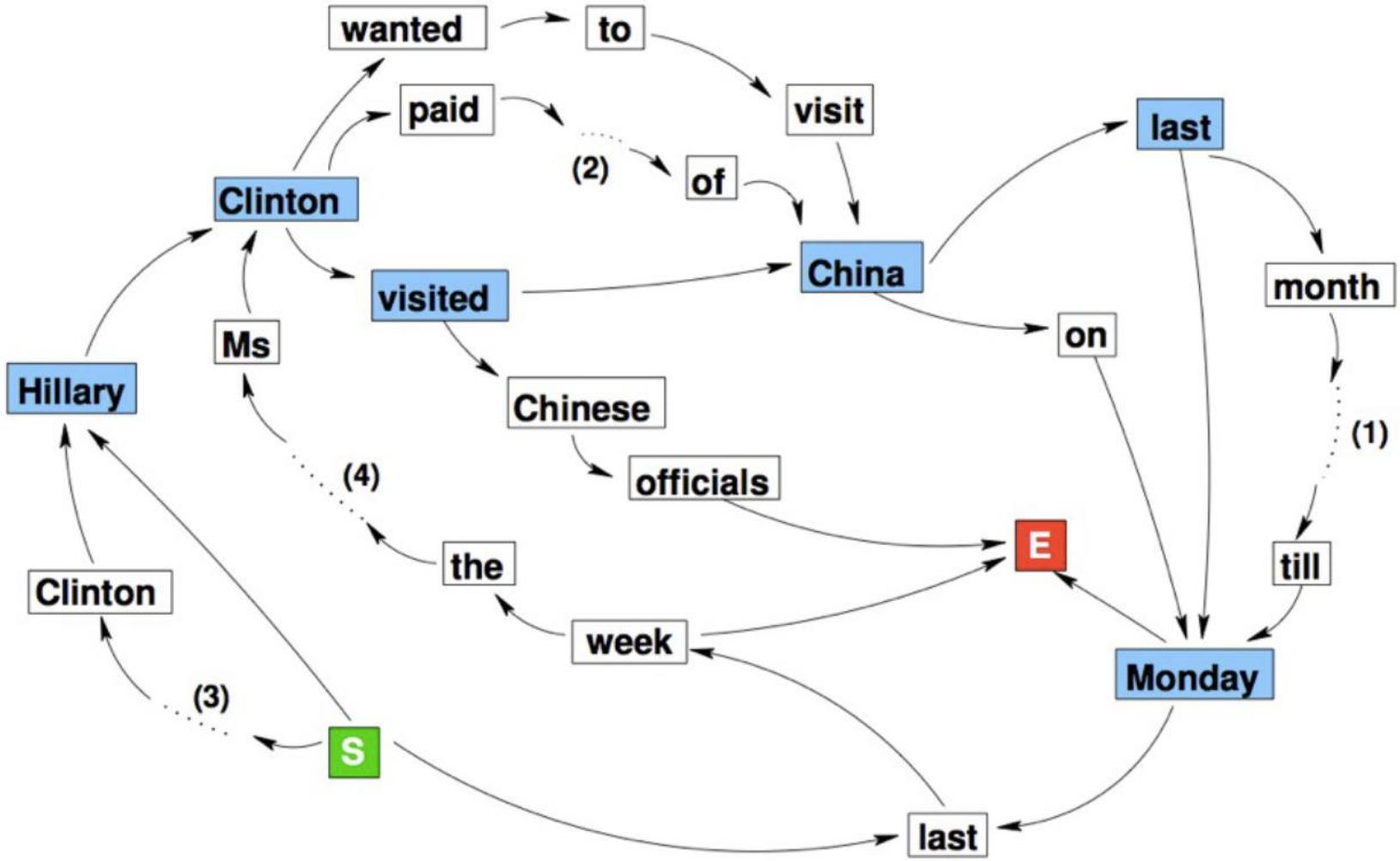
ПОСТОБРАБОТКА

Постобработка

- Возможность сократить реферат
- Сделать его более читаемым

Методы:

- На основе правил (анализ синтаксического разбора)
- Обучение на основе корпуса переформулировок
- Подход на основе «канала с помехами»
- Синтез на основе графового представления (случай многодокументного реферирования)



[Filippova 2011]

Sentence fusion

1 *John Smith, born November 15 1900, studied chemistry and physics at the University of London.*

2 *From 1917 Mr. Smith studied at the University of London and in 1921 he graduated with distinction.*

➡ *Mr. Smith studied chemistry and physics at the University of London from 1917.*

- pieces of related sentences are used to generate a novel sentence
- can be seen as a middle ground between extractive and abstractive summarization
- addresses the incompleteness-redundancy problem

[Filippova]

ОЦЕНКА

Оценка

- сравнение с идеалом
- task-oriented (насколько хорошо помогает решить задачу)

- + связность, удобочитаемость…

Согласие людей - 1

Проблема: низкое согласие людей

Пример:

16 текстов, 3 асессора/текст, по 5 предложений/текст →

182 уникальных предложения.

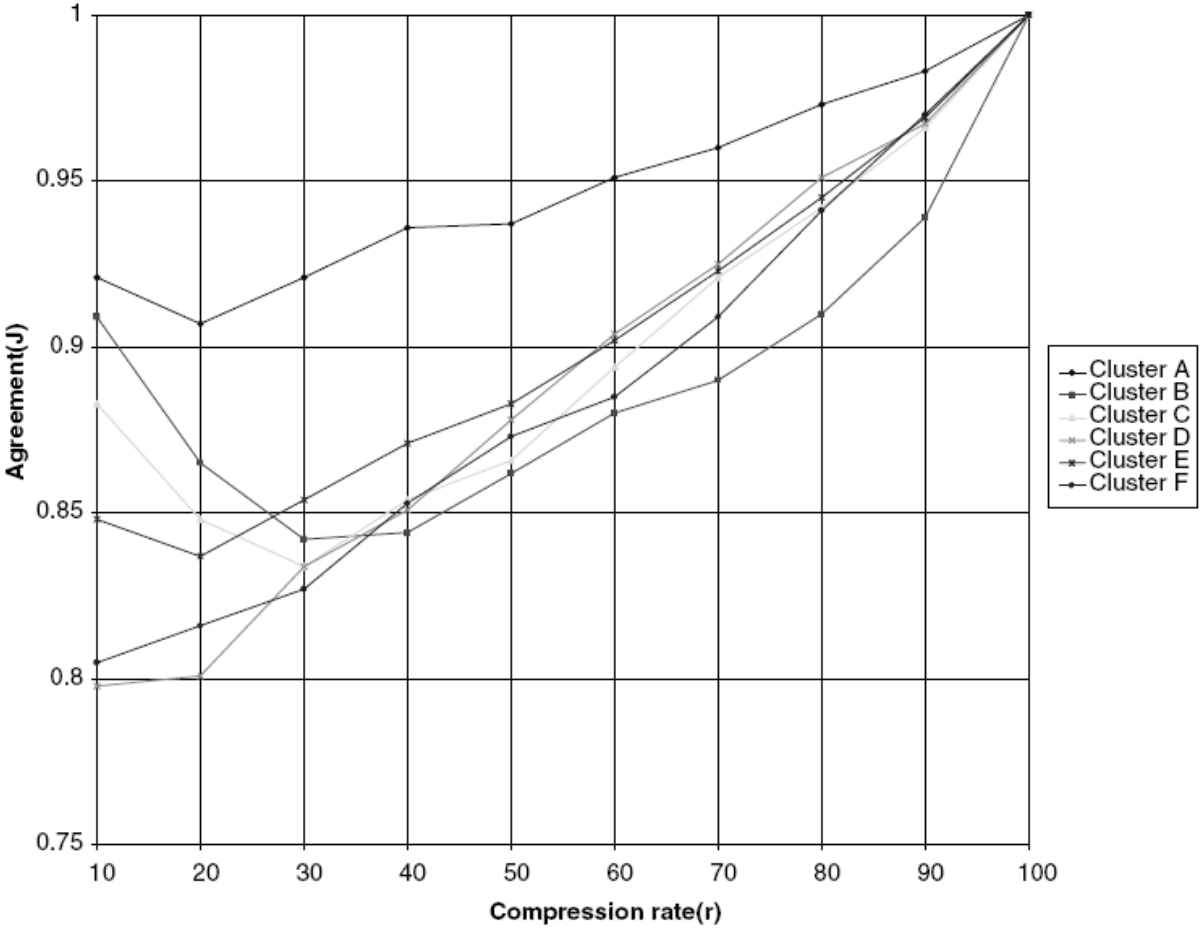
Согласие:

>1: 0,24;

3: 0,05

	# of sent	2 or 3	3	ave	ave 2
T_Imps	12	3	0	0,25	0,00
T_Used	9	5	0	0,56	0,00
T_Weakness	14	1	0	0,07	0,00
T_Awkward	12	2	0	0,17	0,00
T_Use	12	3	0	0,25	0,00
NY_Frantic	12	2	1	0,17	0,08
H_Helm	12	2	0	0,17	0,00
H_Plotting	11	4	1	0,36	0,09
E_Schools	13	0	1	0,00	0,08
H_Tamiflu	12	2	1	0,17	0,08
H_Straight	12	3	0	0,25	0,00
E_Summer	9	4	2	0,44	0,22
E_Decline	11	4	0	0,36	0,00
E_Little	9	3	2	0,33	0,22
H_Treatment	9	4	2	0,44	0,22

Согласие людей - 2



Radev et al. (2004)

Оценка @ DUC

- Generic single document summarization @DUC 2001, 2002
- 100 слов (baseline: первые 100 слов)
- Оценка на основе сравнения с составленным вручную рефератом-моделью: разбивка на предложения с последующим ответом на вопрос «Насколько информация, содержащаяся в предложении модели, передана в оцениваемом реферате?» (6 уровней)
- baseline лучше систем, люди лучше baseline и систем

с 2008 года - TAC

ROUGE (Lin, 2004)

- Recall-Oriented Understudy for Gisting Evaluation
- сравнение реферата с набором «идеальных»
- подсчет общих элементов
- отбор метрик, по которым ранжирование систем коррелирует с ранжированием на основе экспертных оценок (разные для разных задач)

ROUGE package

- ROUGE-N: N-gram co-occurrence statistics is a recall oriented metric

$$\text{ROUGE-n} = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{\text{n-gram} \in S} \text{count}_{\text{match}}(\text{n-gram})}{\sum_{S \in \{\text{Refs}\}} \sum_{\text{n-gram} \in S} \text{count}(\text{n-gram})}$$

[Saggion]

ROUGE package

- ROUGE-L: Based on longest common subsequence
- ROUGE-W: weighted longest common subsequence, favours consecutive matches
- ROUGE-S: Skip-bigram recall metric
- Arbitrary in-sequence bigrams are computed
- ROUGE-SU adds unigrams to ROUGE-S

[Saggion]

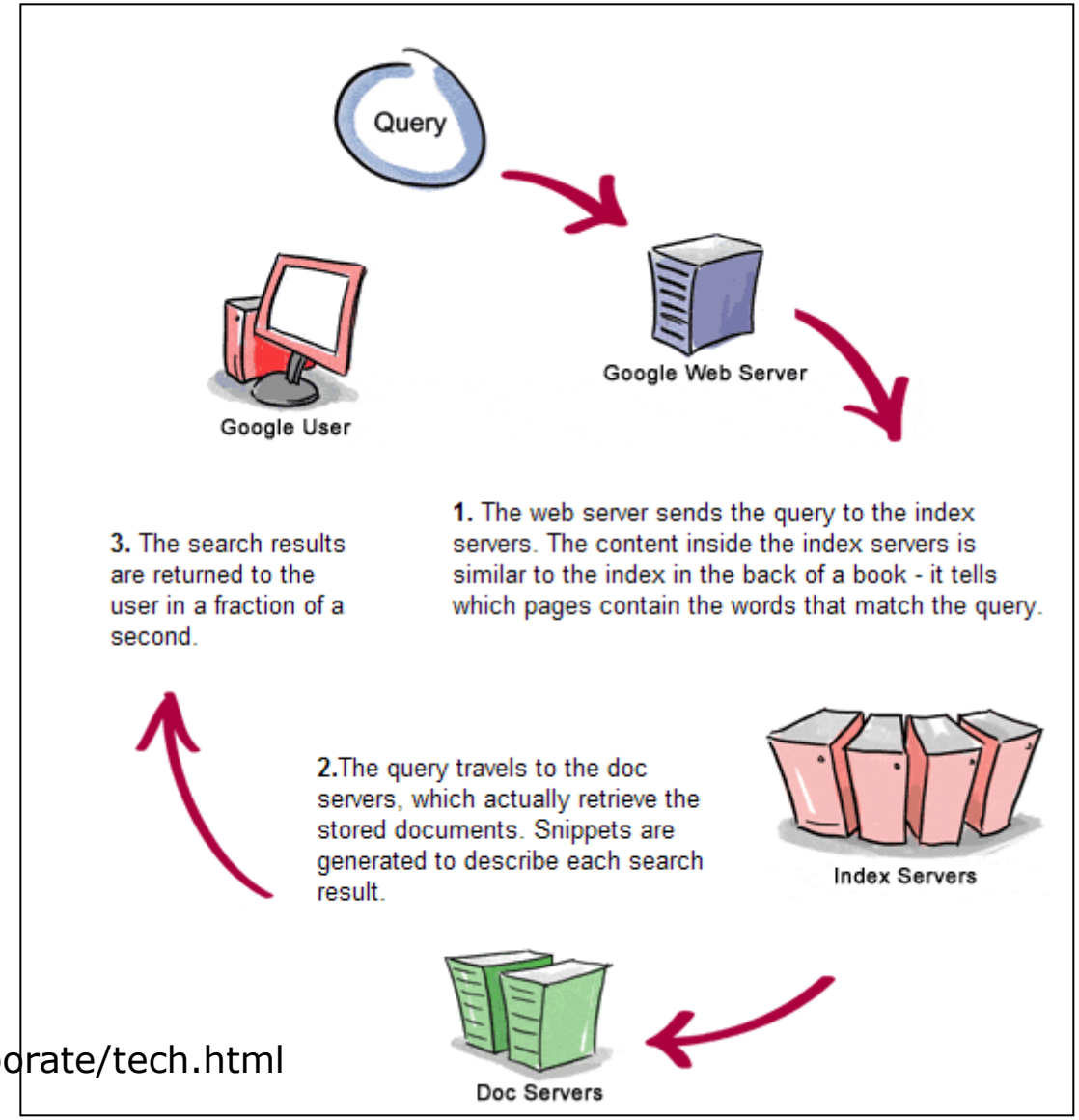
Промежуточные итоги

- простые универсальные методы позволяют получить приемлемый реферат «в среднем»
- повышение качества
  - более глубокий анализ документа
  - больше информации о документе (тема, структура…) → специализация
  - анализ и обработка «сырого» реферата
  - внешний семантические ресурсы

Сниппеты по запросу, Веб-сниппеты

Веб-сниппеты

- Особенности задачи
- Методы
  - рефераты по запросу как частный случай реферирования
  - Веб: использование ссылок
  - Веб: производительность
- Оценка
  - рефераты по запросу
  - инициативы по оценке веб-сниппетов



http://www.google.com/intl/en/corporate/tech.html

Особенности Веб-сниппетов

- документы
  - ссылки, внешние описания
  - структура (title, meta, body, anchor)
  - очень разные (не все являются связным текстом)
- сниппеты для разных запросов (должны быть)
  - индикативными
  - информативными
- UI
  - жесткое ограничение на длину
  - ранжированный список
- спам
- производительность

What are snippets?

Google

Web Images Groups News Scholar more »

sigir Search Advanced Search Preferences

Search: the web pages from Australia

Web

ACM **SIGIR Special Interest Group on Information Retrieval** Home Page
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and distribution ...
www.acm.org/sigs/**sigir**/ - 8k - Cached - Similar pages

SIGIR
The ACM **SIGIR** Conference focuses on research and development in information retrieval. It is the major international forum for the presentation of new ...
www.acm.org/pubs/contents/proceedings/series/**sigir**/ - 22k - Cached - Similar pages
[More results from www.acm.org]

Special Inspector General for Iraq Reconstruction : **SIGIR** Homepage
Welcome to the Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog ...
www.**sigir**.mil/ - 20k - Cached - Similar pages

SIGIR 2006—Seattle
Space Needle **SIGIR** is the major international forum for the presentation of new research results and the demonstration of new systems and techniques in the ...
www.**sigir**2006.org/ - 10k - Cached - Similar pages

[Turpin et al., 2007]

What are snippets?

Google

Web Images Groups News Scholar more »

sigir Search Advanced Search Preferences

Search: the web pages from Australia

Web

ACM **SIGIR Special Interest Group on Information Retrieval** Home Page

"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and distribution ...

www.acm.org/sigs/**sigir**/ - 8k - Cached - Similar pages

SIGIR

The ACM **SIGIR** Conference focuses on research and development in information retrieval. It is the major international forum for the presentation of new ...

www.acm.org/pubs/contents/proceedings/series/**sigir**/ - 22k - Cached - Similar pages

[More results from www.acm.org]

Special Inspector General for Iraq Reconstruction : SIGIR Homepage

Welcome to the Office of the **Special Inspector General for Iraq Reconstruction** (**SIGIR**), a temporary federal agency serving the American public as a watchdog ...

www.**sigir**.mil/ - 20k - Cached - Similar pages

SIGIR 2006—Seattle

Space Needle **SIGIR** is the major international forum for the presentation of new research results and the demonstration of new systems and techniques in the ...

www.**sigir**2006.org/ - 10k - Cached - Similar pages

[Turpin et al., 2007]

InCommonSense

(The|This) Anchor --- ---
- ---- - -- - --- - -------- -
-------- ----- --- ----------

- Выбран наиболее продуктивный шаблон ссылающегося абзаца
- Набор таких абзацев, ссылающихся на страницу, – кандидаты в сниппеты
- Проведена оценка полученных описаний в качестве сниппетов
- построен классификатор хороший/плохой сниппет
- признаки (15): длина, личные местоимения, аббревиатуры, оценочные прилагательные, слова, указывающие на содержание (about, information), повторы слов, оформление (заглавные буквы, пунктуация, глаголы)…

(Amitay&Paris, 2000)

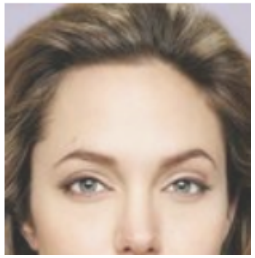
анджелина джоли — 6 млн ответов

Найти

Джоли, **Анджелина** — Википедия

ru.wikipedia.org › **Джоли**, **Анджелина**

Анджели́на Джоли́ Питт (англ. **Angelina Jolie** Pitt, при рождении Войт; род. 4 июня 1975, Лос-Анджелес) — американская актриса, режиссёр и сценарист, фотомодель, посол доброй воли ООН.



Джоли, Анджелина

Материал из Википедии — свободной энциклопедии

В Википедии есть статьи о других людях с фамилией Джоли.

Анджели́на Джоли́ Питт (англ. *Angelina Jolie Pitt*[3], при рождении **Войт**; род. 4 июня 1975, Лос-Анджелес) — американская актриса, режиссёр и сценарист, фотомодель, посол доброй воли ООН.

```
63 <meta property="og:title" content="&#x2605;&#x2605;&#x2605; Гостиница Варшава, Москва, Россия" />
64 <meta property="og:description" content="Трехзвездочная гостиница «Варшава» находится в центре Москвы, в 3
65 <meta property="og:image" content="http://q-ec.bstatic.com/images/hotel/max300/584/5849012.jpg" />
66 <meta property="og:image:width" content="300" />
67 <meta property="og:image:height" content="200" />
```

Производительность

- см. Turpin A. et al. Fast Generation of Result Snippets in Web Search, SIGIR 2007.

- Презентация: http://goanna.cs.rmit.edu.au/~ytsegay/talks/ytsegay_sigir2007.ppt

Эксперименты по оценке

- (Tombros&Sanderson, 1998)
  - простой метод реферирования (title, location, term occurrence, query terms)
  - рефераты по запросу vs. начало документа
  - сравнение:
    - оценки релевантности на рефератах/полных текстах (данные TREC);
    - скорость;
    - обращения к полным документам

Эксперименты по оценке – 2

- (White et al., 2003)
  - простой метод выделения предложений: WebDocSum (title, location, query-biased, text-formatting)
  - единый интерфейс
  - 4 системы: Google +/- WebDocSum, AltaVista +/- WebDocSum
  - task-oriented evaluation:
    - Search for a fact
    - Search for a number of items
    - Decision search
    - Background search
  - Анкеты, мысли вслух, логгирование поведения

РОМИП

- релевантность на сниппетах/полных документах (2005-2007)
  - Асессор оценивает релевантность документа запросу по аннотации этого документа. Сам документ ассессор не увидит. При оценке асессор будет видеть название (первые 100 символов) и аннотацию (300 символов).
- 2008
  - информативность (плохо/хорошо/отлично) («Достаточно ли информации для принятия решения о релевантности?»)
  - читабельность (плохо/хорошо/отлично)
  - оценка релевантности документа