МНОГОФАКТОРНЫЙ МЕТОД ПОСТРОЕНИЯ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ НОВОСТНОГО КЛАСТЕРА

Аспирант ВМК
Алексеев Алексей

ОБРАБОТКА ПОТОКОВ НОВОСТЕЙ

- Новостные сервисы (30-40 тыс. документов в день)
- Кластеризация новостей на одну тему – новостной кластер (вхождения слов)
  - Удаление дубликатов
  - Рубрикация по тематическим рубрикам
  - Автоматическое аннотирование
  - Определение новизны
  - Извлечение информации
- Многие операции выполняются на основе пословного представления

ПРОБЛЕМЫ ПОСЛОВНОГО ПРЕДСТАВЛЕНИЯ

- Одна сущность названа посредством цепочки слов (многословным выражением)
- В кластере используется много разных наименований одной и той же сущности
- Авиабаза США в Киргизии:
  - *база Манас, авиабаза Манас, Манас,*
  - *база в международном аэропорту Манас,*
  - *база США, американская авиабаза*
- **Проблемы:**
  - Определение границ кластера
  - Автоматическое порождение аннотации
  - Определение новизны информации
  - Выделение подкластеров и др.

ПОСТАНОВКА ЗАДАЧИ - 1

- Кластер документов должен соответствовать ситуации или совокупности связанных ситуаций (основная тема кластера)
- В ситуации есть набор участников, которые в кластере:
  - Могут быть выражены не только словами, но и словосочетаниями
  - Каждый участник может выражаться совокупностью разных выражений
- Тестирование показывает, что качество кластеров оценивается около **70**% (Добров Б.В,Павлов А.С. 2010), т.е. основные участники в среднем представлены правильно
- **Предположение:** более точное выявление основных участников ситуации поможет более правильно осуществлять разные операции с кластерами

ПОСТАНОВКА ЗАДАЧИ - 2

- **Входные данные:**
  Кластер тематически близких документов, собранных на основе пословной модели представления документов

- **Задача:**
  Построить тематическое представление входного новостного кластера – совокупность основных участников ситуации и отношений между ними

- **Применение:**
  - Добавление в пословное представление документа выделенных структур
  - Использование усложненного представления документов кластера для улучшения качества операций с кластером

ПЛАН ДОКЛАДА

- **Природа вариативности и обзор существующих методов**

- Тематическое представление текста
- Метод выделения структур, описывающих участников ситуации
- Тестирование подхода в автоматическом аннотировании
- Заключение

ПРИРОДА ВОЗНИКНОВЕНИЯ ВАРИАТИВНОСТИ - 1

- **Цель использования:**
  - ***Референция*** (отнесенность языкового выражения к одному и тому же объекту действительности)

*3 февраля **президент Киргизии Курманбек Бакиев** заявил о **решении правительства** прекратить деятельность авиабазы на территории республики… **Президент** не стал скрывать, что экономические резоны стали главной причиной побудившей правительство страны принять такое **решение**.*

  - ***Перефразирование*** (изменение текста без изменения смысла - рерайтинг)

*Судьбу авиабазы США в "Манасе" **решит** парламент Киргизии. Парламент Киргизии в четверг **примет окончательное решение** о судьбе авиабазы США.*

ПРИРОДА ВОЗНИКНОВЕНИЯ ВАРИАТИВНОСТИ - 2

- **Привязка к контексту:**
  - ***Общеизвестно*** (*Киргизия – Кыргызстан*)

*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*

  - ***Выводится из контекста***

В *декабре 2006 года 46-летний водитель топливозаправщика киргизской фирмы, занимающейся обслуживанием аэропорта "Манас", Александр Иванов, был расстрелян в упор **охранником авиабазы** Закари Хатфилдом на КПП при въезде на перрон аэропорта"… **Американский военный**, несмотря на неоднократные требования киргизского МИДа, также был тайно вывезен с территории страны и до сих пор не предстал перед судом.*

УСЛОЖНЕНИЕ ПОСЛОВНОГО ПРЕДСТАВЛЕНИЯ
ЛЕКСИЧЕСКИЕ ЦЕПОЧКИ - 1

- Использование информации об объектах и связях между ними, описанной в предопределенных ресурсах
- **Популярные подходы:**
  - *Английский язык*: ***Wordnet***

    Barzilay R., Elhadad M., 1999
  - *Русский язык*: ***РуТез***

    Лукашевич Н.В., Добров Б.В., 2009

Лексические цепочки в форме ***тематических узлов*** - к одному выделенному элементу относятся все другие элементы лексической цепочки

УСЛОЖНЕНИЕ ПОСЛОВНОГО ПРЕДСТАВЛЕНИЯ
ЛЕКСИЧЕСКИЕ ЦЕПОЧКИ - 2

Пример основных тематических узлов на основе РуТез:

ТЕННИСНЫЙ КОРТ	**14**
ТЕННИС	**12**
АВСТРИЙЦЫ	**12**
АВСТРИЯ	**6**
КИПРИОТЫ	**16**
КИПР	**11**
ХОРВАТЫ	**10**
СЕТ (ПАРТИЯ В ТЕННИСЕ)	**6**
ИГРОВАЯ ПАРТИЯ	**5**
ЧЕТВЕРТЬФИНАЛ	**10**
ПОЛУФИНАЛ	**29**
ПОЛУФИНАЛИСТ	**2**

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

УСЛОЖНЕНИЕ ПОСЛОВНОГО ПРЕДСТАВЛЕНИЯ
РЕФЕРЕНЦИАЛЬНЫЕ ЦЕПОЧКИ

- Выявление различных упоминаний (отсылок) к одной и той же ***реальной*** сущности
- Учет различных видов отсылок, в том числе местоимений

Пример:

Президент РФ *положительно оценил отчет* ***премьера*** *в Госдуме*

За выступлением ***премьера*** *в Думе наблюдал* ***Владимир Путин***. ***Он*** *просмотрел большую часть выступления* ***Дмитрия Медведева*** *и в целом дал позитивную оценку отчёту главы правительства, сообщил журналистам пресс-секретарь* ***Президента*** *Дмитрий Песков.*

- Узкая специализация: выявление строго-определенного спектра сущностей – люди, организации, места и т.д.

Latent Dirichlet Allocation (LDA)

University of California, USA, 2003
David M. Blei, Andrew Y. Ng, Michael I. Jordan

- Вероятностная модель, разбивающая слова на тематические группы
- Двухуровневое представление:
  - Распределение Topics-VS-Documents (TvsD)
  - Распределение Words-VS-Topics (WvsT)
- Наблюдаемый документ – автоматический семплинг из указанных распределений:
  I. Выбор топика из TvsD
  II. Выбор слова из WvsT
- **Проблемы:**
  - Вероятностный результат
  - Сложность интерпретации результатов работы алгоритма

Latent Dirichlet Allocation (LDA) ПРИМЕР

"Arts" "Budgets" "Children" "Education"

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

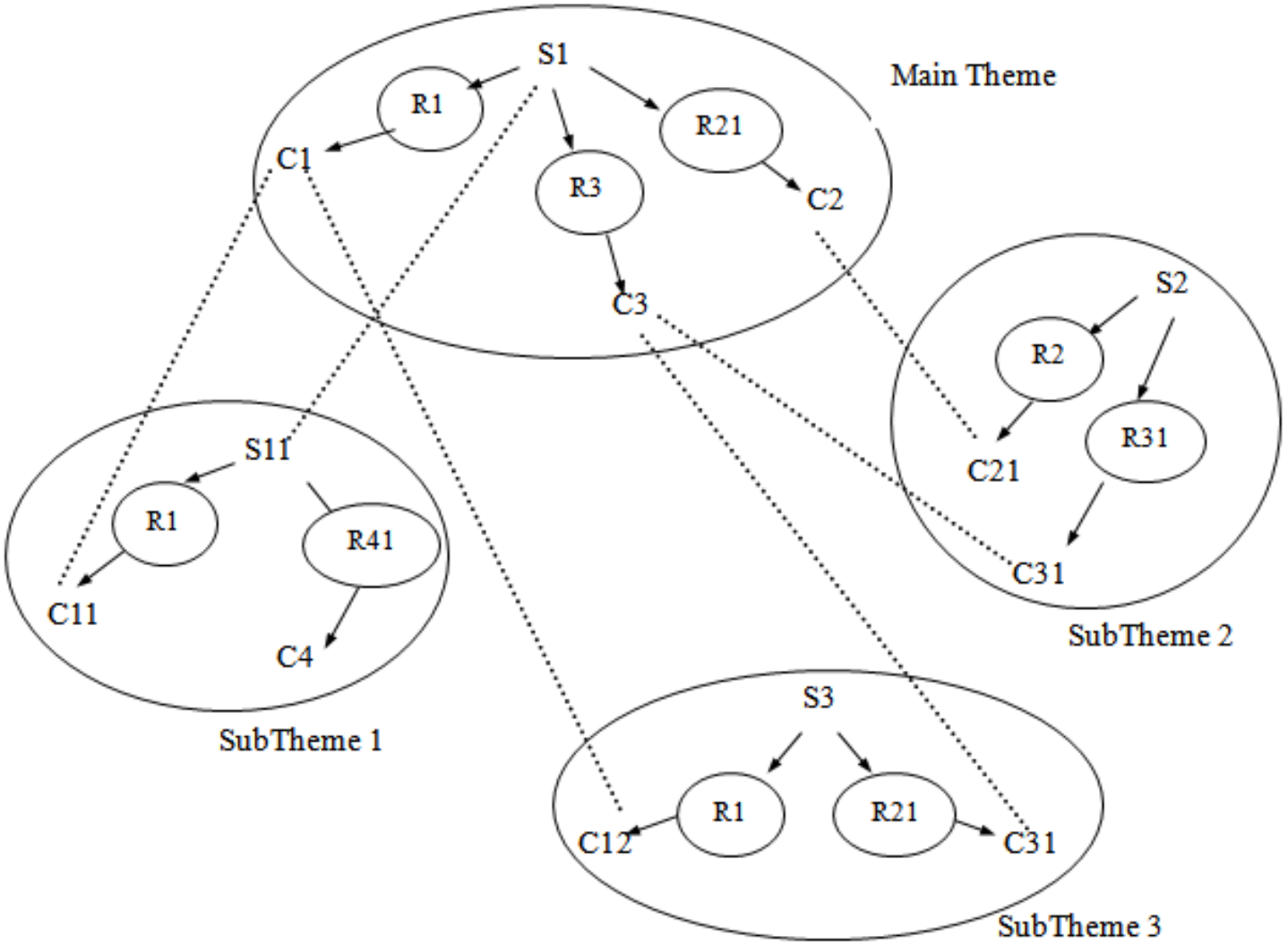
ПЛАН ДОКЛАДА

- Природа вариативности и обзор существующих методов
  - **Тематическое представление текста**
- Метод выделения структур, описывающих участников ситуации
- Тестирование подхода в автоматическом аннотировании
- Заключение

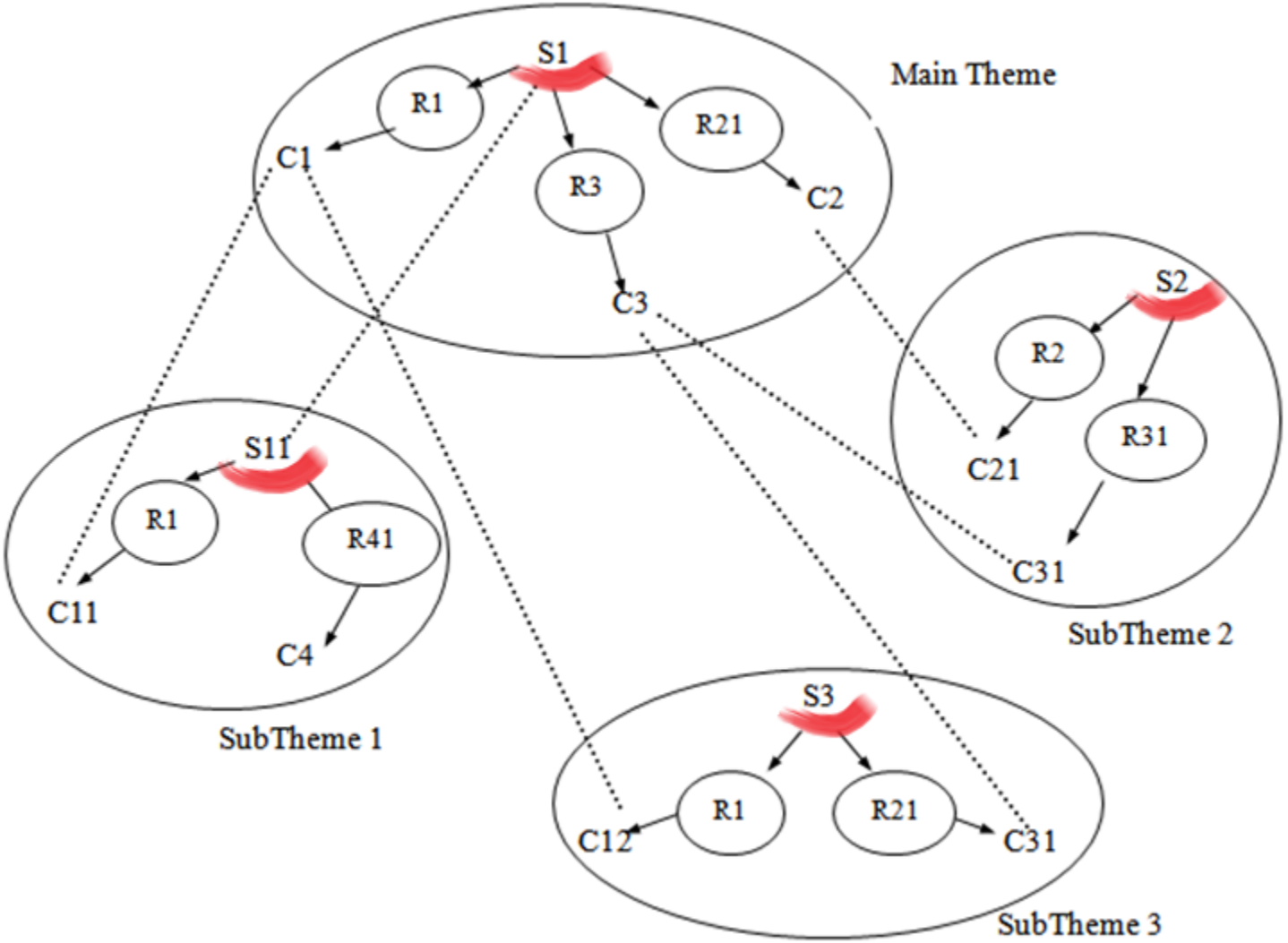
ГЛОБАЛЬНАЯ СВЯЗНОСТЬ ТЕКСТА

- **Van Dijk и гипотеза глобальной связности (1985)**
- Связный текст имеет одну главную тему и эта тема может быть выражена как пропозиция
- Тема целого текста раскрывается в тексте посредством локальных тем
- Каждое предложение текста соответствует некоторой теме текста
- Механизм глобальной связности позволяет контролировать локальный связки и переходы

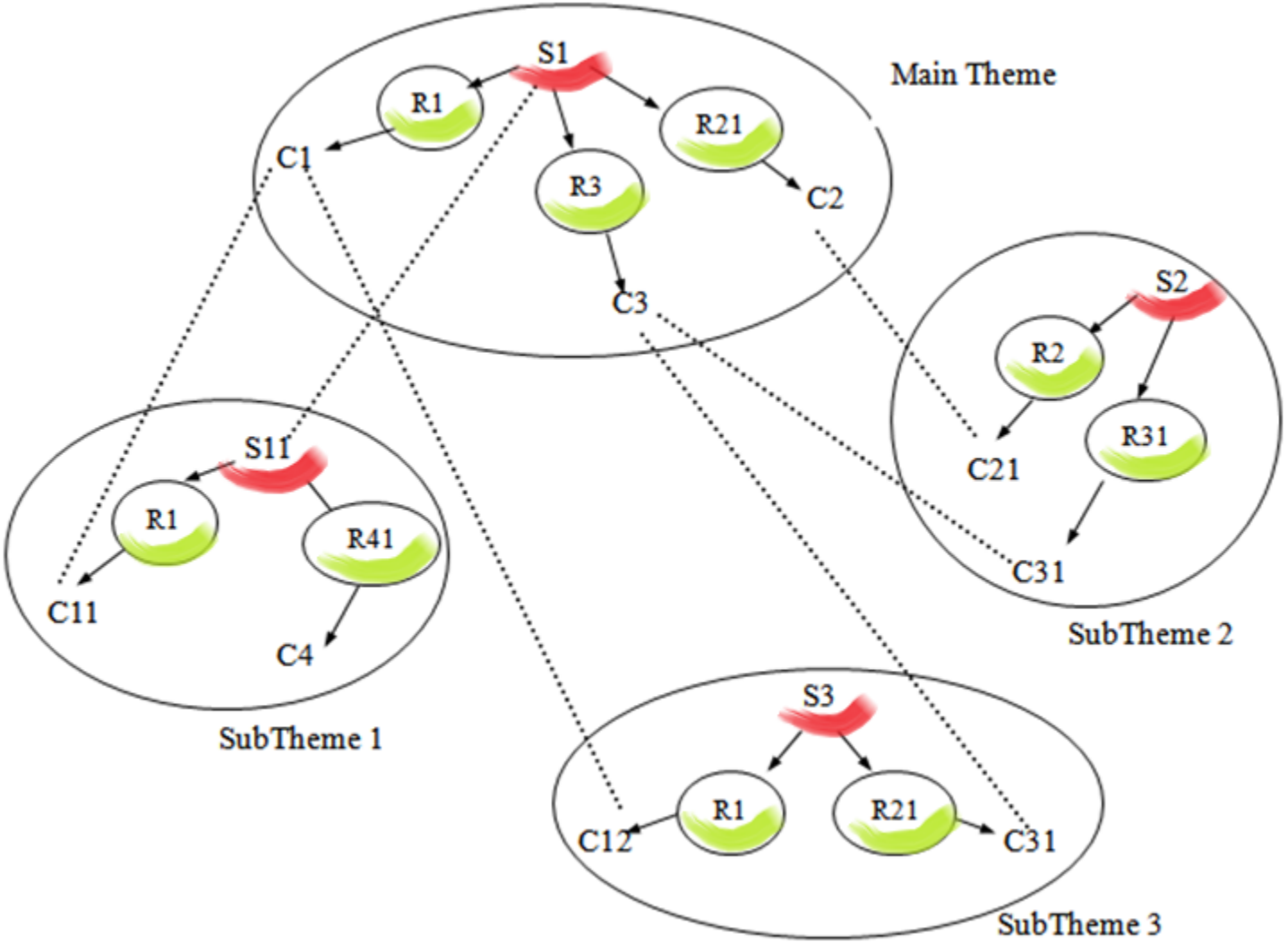
ГЛАВНАЯ ТЕМА И ЕЕ УТОЧНЕНИЕ В КОНКРЕТНЫХ ПРЕДЛОЖЕНИЯХ



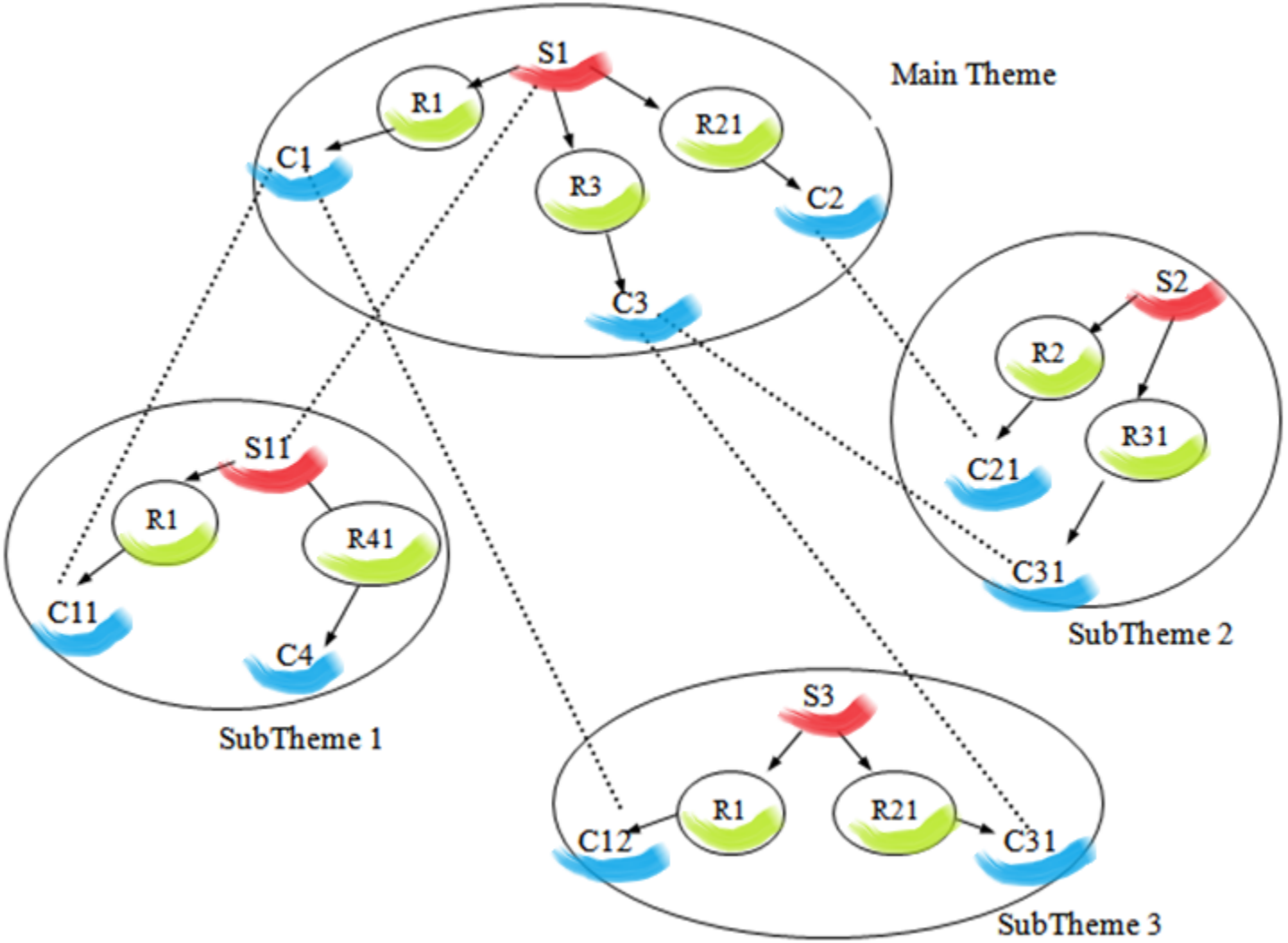
ГЛАВНАЯ ТЕМА И ЕЕ УТОЧНЕНИЕ В КОНКРЕТНЫХ ПРЕДЛОЖЕНИЯХ



ГЛАВНАЯ ТЕМА И ЕЕ УТОЧНЕНИЕ В КОНКРЕТНЫХ ПРЕДЛОЖЕНИЯХ



ГЛАВНАЯ ТЕМА И ЕЕ УТОЧНЕНИЕ В КОНКРЕТНЫХ ПРЕДЛОЖЕНИЯХ



ЛЕКСИЧЕСКАЯ СВЯЗНОСТЬ vs. ГЛОБАЛЬНАЯ СВЯЗНОСТЬ

- Связный текст обладает лексической связностью: лексические и семантические повторы
- Лексическая связность – инструмент глобальной связности
- Чем больше две сущности упоминаются в одних и тех же предложениях текста, тем более важно отношение между ними для содержания текста
- Если сущности часто упоминаются, но мало встречаются в одних и тех же предложениях текста, то возможно они связаны между собой по смыслу (синоним, род-вид, часть-целое)
- Элементы тематического представления - имеют внутреннюю структуру тематического узла (ТУ): главное выражение и относящиеся к нему объекты

ПРОВЕРКА ГИПОТЕЗЫ ВСТРЕЧАЕМОСТИ В ПРЕДЛОЖЕНИЯХ - 1

- Проверка предположений была произведена с помощью Тезауруса русского языка РуТез
- В качестве правильных примеров (связанных языковых выражений) рассматривались выражения имеющие связь по Тезаурусу
- Различные типы связи рассматривались отдельно (***синонимия; род-вид; часть-целое***)
- Две группы по частям речи:

СУЩ. + СУЩ. **#** **ПРИЛ. + СУЩ.**

- Для каждой пары объектов вычислялись количество вхождений в одни и те же предложения (***Fsegm***) и в соседние (***Fsent***)

ПРОВЕРКА ГИПОТЕЗЫ ВСТРЕЧАЕМОСТИ В ПРЕДЛОЖЕНИЯХ - 2

Тип связи	Fsegm / Fsent	Число пар
Синонимы (СУЩ + СУЩ)	**0.309**	31
Синонимы (ПРИЛ + СУЩ)	**0.491**	53
Род – Вид (СУЩ + СУЩ)	1.130	88
Род – Вид (ПРИЛ + СУЩ)	1.471	28
Часть – Целое (СУЩ + СУЩ)	0.779	58
Часть – Целое (ПРИЛ + СУЩ)	1.580	29
Без связи по Тезаурусу	**1.440**	21483

Вывод: превышение Fsent по отношению к Fsegm свидетельствует о связи между объектами

НОВОСТНЫЕ КЛАСТЕРЫ И СВОЙСТВА СВЯЗНОГО ТЕКСТА

- **Кластер** – не является связным текстом,
  - но имеет тему кластера
  - статистические особенности усиливаются
- **Извлечение**
  - Многословных выражений,
  - Структур, описывающих основных участников ситуации (тематических узлов)
- **Пример:** Новостной кластер от 03.02.2007
  - Тема: Отставка президента алмазодобывающей компании АК «Алроса» Александра Ничипорука
  - 12 новостных документов

ПЛАН ДОКЛАДА

- Природа вариативности и обзор существующих методов
- Тематическое представление текста
  - **Метод выделения структур, описывающих участников ситуации**
- Тестирование подхода в автоматическом аннотировании
- Заключение

ПРИНЦИПЫ ФОРМИРОВАНИЯ ТЕМАТИЧЕСКИХ УЗЛОВ

- Структуры, соответствующие участникам ситуации, собираются в форме тематических узлов:
  - центр и элементы ("похожие" на центр слова)
  - элементы могут принадлежать нескольким (двум) тематическим узлам
- Элементы тематического узла могут быть "похожи" на центр по нескольким параметрам
- Если два выражения часто употребляются в одном предложении, они не могут оказаться в одном тематическом узле - ограничивающий фактор
- Предварительная сборка многословных выражений

ИЗВЛЕЧЕНИЕ МНОГОСЛОВНЫХ ВЫРАЖЕНИЙ

- Рассматривается как необходимый этап для построения тематических узлов
- Превышение встречаемости слов непосредственно рядом друг с другом по сравнению с раздельной встречаемостью во фрагментах предложений:

$$Near > 2 * (Av + NotN)$$

- Дополнительные ограничения по частотности встречаемости слов рядом друг с другом
- **Примеры:** *Президент Алроса; президент компании; владелец компании; контрольный пакет акций; акции компании; пакет акций*

АЛГОРИТМ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

- Разные признаки сходства:
  - **Контекстно-независимые признаки** (формальное сходство, тезаурус)
  - **Контекстно-зависимые признаки** («строгие» контексты, статистика контекстов)
- 4 типа контекстов: через глагол (**AV**); рядом (**Near**); не рядом (**NN**); в соседних предложениях (**NS**)
- Итеративный алгоритм. Одна итерация = объединение одной пары объектов
- **Управляющее правило:**

$$NS >= 2 * (AV + Near + NN)$$

КОНТЕКСТНО-НЕЗАВИСИМЫЕ ПРИЗНАКИ

A. **Формальное сходство объектов (BS)**

- Простая метрика – одинаковые начала слов
- Пример: *Руководитель – Руководство*
- Вес признака: [0, 1]

$$BS = 1.0 - 0.1 * (\{ Число\_слов\_с\_отличными\_началами \})$$

B. **Наличие связи в тезаурусе (TS)**

- Рассмотрение цепочек отношений между объектами (использование вывода связей)
- Пример: *Руководитель – Глава*
- Вес признака: [0, 1]

$$TS = 1.0 - 0.2 * (\{ Длина\_пути\_по\_отношениям\_тезауруса \})$$

КОНТЕКСТНО-ЗАВИСИМЫЕ ПРИЗНАКИ

A. **Частота появления в соседних предложениях (NSF)**

$$NSF = Min\left[1, \frac{C}{Avg(C)}\right] \qquad C = NS - 2 * \left(AcrossVerb + Near + NotNear\right)$$

B. **Анализ «строгих» контекстов (SC)**

- 4-граммы: по 2 слова влево и вправо
- Значение: [0, 1] (относительно «лучшей» пары)

C. **Скалярные произведения контекстов (SPS)**

- Косинусная мера угла между векторами контекстов
- Сумма по характеристикам внутри предложения (AV + Near + NotNear)

ПРИМЕР РАНЖИРОВАНИЯ ПАР

Features / Pairs	Context-independent		Context-dependent			SCORE
	BS	TS	NSF	SC	SPS	
Президент России – Президент РФ	0.90	1.00	0.00	0.50	0.68	**2.74**
Инвестгруппа– Инвестиционная группа	0.90	1.00	0.40	0.00	0.63	**2.42**
ГМК Норильский никель – Норильский никель	1.00	1.00	0.40	0.00	0.21	**2.31**
Российская Федерация – Россия	0.90	1.00	0.00	0.00	0.51	**2.15**
Отставка – Отставка с должности	0.90	1.00	0.40	0.00	0.00	**2.10**

ПРИМЕР СБОРКИ УЗЛА

- **Итерация 7:** (***Отставка)*** ← *(**Отставка с должности**)*
- **Итерация 33:** (***Отставка,*** *Отставка с должности)* ← *(**Уход в отставку**)*
- **Итерация 44:** (***Отставка,*** *Отставка с должности, Уход в отставку)* ← *(**Отставка президента**)*
- **Итерация 61:** (***Уход с поста)*** ← *(**Уход в отставку**)*
- **Итерация 62:** (***Отставка,*** *Отставка с должности, Уход в отставку, Отставка президента)* ← *(**Уход с поста,** Уход в отставку)*
- **Итерация 102:** (***Отставка,*** *Отставка с должности, Уход в отставку, Отставка президента, Уход с поста)* ← *(**Пост**)*
- **Итерация 103:** (***Пост,*** *Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста)* ← *(**Должность**)*
- **Итерация 104:** (***Пост,*** *Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста, Должность)* ← *(**Уход**)*

РЕЗУЛЬТАТЫ РАБОТЫ АЛГОРИТМА НА КЛАСТЕРЕ ПРИМЕРЕ

- ***Пост:*** *уход с поста; должность; уход; отставка; отставка с должности; уход в отставку; отставка президента*
- ***Алроса:*** *президент Алроса; АК Алроса*
- ***Компания****: акция компании; владелец компании; объединение компаний; акция; акционер компании; владелец; пакет акций; состав владельцев; контрольный пакет акций; контрольный пакет; владение*
- ***Ничипорук****: Александр Ничипорук*
- ***Якутия****: президент Якутии; якутский; якутский президент*

ПЛАН ДОКЛАДА

- Природа вариативности и обзор существующих методов
- Тематическое представление текста
- Метод выделения структур, описывающих участников ситуации
  - **Тестирование подхода в автоматическом аннотировании**
- Заключение

ПРИМЕНЕНИЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

- *Тематическое представление* – дополнительная информация об устройстве новостного кластера

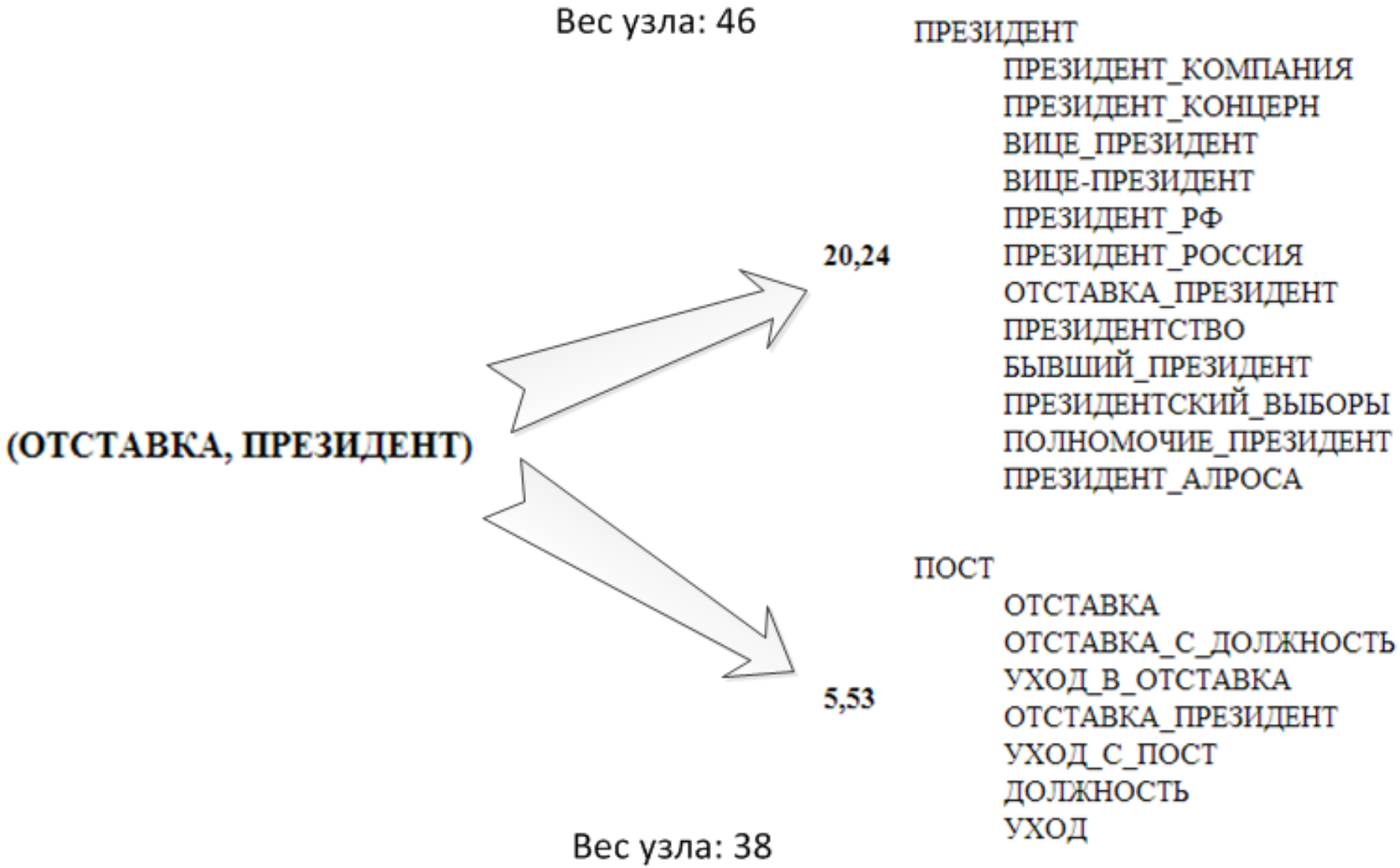
Слова → ***Объекты*** *(слова + МВ)* → ***Тематические узлы***

- Вес тематического узла:

$$weight(ТУ) = \sum_{ЭЛЕМЕНТ\_ТУ_i \in ТУ} freq(ЭЛЕМЕНТ\_ТУ_i)$$

- Вес объекта разбивается на веса в соответствующих тематических узлах (пропорционально силе связи с центром узла)

ПРИМЕНЕНИЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ - ПРИМЕР



ТЕМАТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ И АННОТИРОВАНИЕ

- Известные метода аннотирования:
  - Maximal Marginal Relevance (MMR)
  - SumBasic
  - Аннотирование на основе тем-го представления РуТез
- Специализированные методы аннотирования
  - Учет основных тематических узлов
  - Учет основных связей тематических узлов
- Учет **IDF**:

$$MWE\_IDF = \ln\left(\frac{DOC\_COUNT}{[\prod_{w_i \in MWE} Freq(w_i)]/[DOC\_COUNT]^{N-1}}\right)$$

Maximal Marginal Relevance (MMR)

- Известный метод для запрос-ориентированного аннотирования (Carbonell, Goldstein, 1998)
- Итеративный метод
- Ранжирование предложений-кандидатов:
  - ✓ Максимизировать сходство с запросом
  - ✓ Минимизировать сходство с уже отобранными в аннотацию предложениями

Пусть: Q – запрос к системе, **S** – множество предложений кандидатов, **s** – рассматриваемое предложение кандидат, **E** – множество выбранных предложений. Тогда:

$$MMR = \arg\max_{s \in S} \left[\lambda \cdot Sim_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} Sim_2(s, s_j) \right]$$

- ***Добавление тематического представления:*** вектора кластера и предложений строятся по тематическим узлам

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:

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

n – число вхождений слова

N – общее число токенов

- 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

SumBasic
+ ТЕМАТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ

$$p_{new}(w_i) = p_{old}(w_i) * p_{old}(w_i)$$

- Переход от вероятностей отдельных слов к вероятностям тематических узлов:

$$p(w_i) = \{p(T_1), p(T_2), \ldots p(T_K)\}$$ **K** – число ТУ

- В расчете веса предложения участвует тематический узел с бОльшей вероятностью:

$$weight(w_i) = \max(\{p(T_1), p(T_2), \ldots p(T_K)\})$$

- Понижение веса для отобранного предложения происходит для тех же узлов, что участвовали при расчете веса предложения

АННОТИРОВАНИЕ НА ОСНОВЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ПО ТЕЗАУРУСУ РУТЕЗ

- «*Автоматическое аннотирование новостных кластеров на основе тематического представления*» (Лукашевич Н.В., Добров Б.В, 2009)
- **Пусть**: *ТУ_ТЕЗ* – тематический узел на основе тезауруса РуТез
  ИС – именованная сущность
  * *OLD* – упомянутый ТУ_ТЕЗ или ИС
  * *NEW* – новый ТУ_ТУЗ или ИС
- ***Тогда на каждой итерации алгоритма выбирается предложение:***

1. $S_i \supset \{ТУ\_ТЕЗ\_OLD \vee ИС\_OLD\}$
2. $S_i \supset \{ТУ\_ТЕЗ\_NEW \vee ИС\_NEW\}$
3. $\max\left(\sum_{ТУ\_ТЕЗ \in S_i} weight(ТУ\_ТЕЗ) + \sum_{ИС \in S_i} weight(ИС) \right)$

СОБСТВЕННЫЕ АЛГОРИТМЫ АННОТИРОВАНИЯ НА ОСНОВЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

Пусть: *ТУ* – тематический узел на основе предложенного алгоритма

ТУ_REL – пара ТУ, **_NEW* – новый ТУ или ТУ_REL

Предлагается два алгоритма аннотирования:

I. **Отбор по тематическим узлам (*OurSummary_Nodes*)**

- $$s_i \Rightarrow \max\left(\sum_{TY\_NEW_j \in s_i,\, i=1..3}^{desc\,weight(TY\_NEW_j)} weight(TY\_NEW_j)\right)$$

II. **Отбор по связям ТУ (*OurSummary_Relations*)**

- weight(ТУ_REL) – число вхождений в одни и те же предложения
- $$s_i \Rightarrow \max_{TY\_REL\_NEW_j \in Cluster}\left(weight(TY\_REL\_NEW_j)\right)$$
- $$s_i \Rightarrow \max\left(\sum_{TY\_REL\_NEW_j \in s_i} weight(TY\_REL\_NEW_j)\right)$$

АЛГОРИТМ ТЕСТИРОВАНИЯ

- Оценка результата нетривиальна: высокая степень субъективности и низкая согласованность экспертов
- 11 новостных кластеров, 2-4 ручные аннотации к каждому
- Оценка содержания аннотаций методом ROUGE
- Различные алгоритмы аннотирования в различных модификациях:
  - Maximal Marginal Relevance (MMR, 4 модификации)
  - SumBasic (2 модификации)
  - Собственный алгоритм аннотирования (4 модификации)
  - Аннотирование на основе тем-го представления РуТез
- Модификации алгоритмов как на основе пословной модели, так и на основе тематического представления

ПРИМЕР АННОТАЦИИ

Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.

1. 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.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

Президент "АЛРОСА“ Александр Ничипорук уходит со своего поста.

1. 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.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

*Президент "АЛРОСА“ **Александр Ничипорук** уходит со своего поста.*

*1. 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.*

*2. **Александр Ничипорук** покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.*

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

Президент "АЛРОСА“ ***Александр Ничипорук*** *уходит со своего поста.*

1. 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.*

2. ***Александр Ничипорук*** *покидает пост* ***президента компании*** *АЛРОСА, на которую приходится четверть мировой добычи алмазов.*

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

Президент "АЛРОСА“ Александр Ничипорук уходит со своего поста.

1. 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.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

Президент "АЛРОСА“ Александр Ничипорук уходит со своего поста.

1. 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.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

МЕТОД ROUGE

- ROUGE - Recall-Oriented Understudy for Gisting Evaluation (Chin-Yew Lin, 2004)
- **Основная идея:** сопоставление порожденных аннотаций с ручными (составленными экспертами)
- Различные метрики сравнения: N-граммы, Longest Common Substring и их модификации

$$ROUGE-N=\frac{\sum_{S\in\{\text{Reference Summaries}\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{\text{Reference Summaries}\}}\sum_{gram_n\in S}Count(gram_n)}$$

РЕЗУЛЬТАТЫ

Метод	1	2	L	S	SU	Avg
MMR + Groups	0,58538 (1)	0,37498 (2)	0,55918 (2)	0,31262 (1)	0,32351 (1)	**1,4**
MMR_WithIDF	0,57996 (3)	0,3833 (1)	0,55891 (3)	0,29818 (2)	0,31003 (2)	**2,2**
MMR_WithoutIDF	0,58128 (2)	0,35748 (3)	0,56248 (1)	0,27601 (3)	0,28815 (3)	**2,4**
OurSummary_Nodes	0,54595 (5)	0,32606 (5)	0,52076 (5)	0,27397 (4)	0,285 (4)	**4,6**
OurSummary_Nodes_WithIDF	0,55526 (4)	0,32374 (6)	0,52925 (4)	0,26181 (7)	0,2738 (7)	**5,6**
OurSummary_Relations	0,54307 (6)	0,30604 (7)	0,51626 (6)	0,26375 (5)	0,27526 (5)	**5,8**
ThematicLines	0,52831 (7)	0,3319 (4)	0,50912 (7)	0,26312 (6)	0,27431 (6)	**6,0**
SumBasic + Groups	0,51884 (9)	0,24591 (10)	0,48974 (9)	0,25057 (8)	0,26142 (8)	**8,8**
MMR_WithIDF + Groups	0,45376 (10)	0,27439 (8)	0,43607 (10)	0,23926 (9)	0,24899 (9)	**9,2**
SumBasic	0,52753 (8)	0,23993 (11)	0,49326 (8)	0,23434 (10)	0,24628 (10)	**9,4**
OurSummary_Relations_WithIDF	0,45172 (11)	0,25309 (9)	0,4333 (11)	0,17712 (11)	0,1882 (11)	**10,6**

РЕЗУЛЬТАТЫ

Метод	1	2	L	S	SU	Avg
MMR + Groups	0,58538 (1)	0,37498 (2)	0,55918 (2)	0,31262 (1)	0,32351 (1)	**1,4**
MMR_WithIDF	0,57996 (3)	0,3833 (1)	0,55891 (3)	0,29818 (2)	0,31003 (2)	**2,2**
MMR_WithoutIDF	0,58128 (2)	0,35748 (3)	0,56248 (1)	0,27601 (3)	0,28815 (3)	**2,4**
OurSummary_Nodes	0,54595 (5)	0,32606 (5)	0,52076 (5)	0,27397 (4)	0,285 (4)	**4,6**
OurSummary_Nodes_WithIDF	0,55526 (4)	0,32374 (6)	0,52925 (4)	0,26181 (7)	0,2738 (7)	**5,6**
OurSummary_Relations	0,54307 (6)	0,30604 (7)	0,51626 (6)	0,26375 (5)	0,27526 (5)	**5,8**
ThematicLines	0,52831 (7)	0,3319 (4)	0,50912 (7)	0,26312 (6)	0,27431 (6)	**6,0**
SumBasic + Groups	0,51884 (9)	0,24591 (10)	0,48974 (9)	0,25057 (8)	0,26142 (8)	**8,8**
MMR_WithIDF + Groups	0,45376 (10)	0,27439 (8)	0,43607 (10)	0,23926 (9)	0,24899 (9)	**9,2**
SumBasic	0,52753 (8)	0,23993 (11)	0,49326 (8)	0,23434 (10)	0,24628 (10)	**9,4**
OurSummary_Relations_WithIDF	0,45172 (11)	0,25309 (9)	0,4333 (11)	0,17712 (11)	0,1882 (11)	**10,6**

РЕЗУЛЬТАТЫ

Метод	1	2	L	S	SU	Avg
MMR + Groups	0,58538 (1)	0,37498 (2)	0,55918 (2)	0,31262 (1)	0,32351 (1)	**1,4**
MMR_WithIDF	0,57996 (3)	0,3833 (1)	0,55891 (3)	0,29818 (2)	0,31003 (2)	**2,2**
MMR_WithoutIDF	0,58128 (2)	0,35748 (3)	0,56248 (1)	0,27601 (3)	0,28815 (3)	**2,4**
OurSummary_Nodes	0,54595 (5)	0,32606 (5)	0,52076 (5)	0,27397 (4)	0,285 (4)	**4,6**
OurSummary_Nodes_WithIDF	0,55526 (4)	0,32374 (6)	0,52925 (4)	0,26181 (7)	0,2738 (7)	**5,6**
OurSummary_Relations	0,54307 (6)	0,30604 (7)	0,51626 (6)	0,26375 (5)	0,27526 (5)	**5,8**
ThematicLines	0,52831 (7)	0,3319 (4)	0,50912 (7)	0,26312 (6)	0,27431 (6)	**6,0**
SumBasic + Groups	0,51884 (9)	0,24591 (10)	0,48974 (9)	0,25057 (8)	0,26142 (8)	**8,8**
MMR_WithIDF + Groups	0,45376 (10)	0,27439 (8)	0,43607 (10)	0,23926 (9)	0,24899 (9)	**9,2**
SumBasic	0,52753 (8)	0,23993 (11)	0,49326 (8)	0,23434 (10)	0,24628 (10)	**9,4**
OurSummary_Relations_WithIDF	0,45172 (11)	0,25309 (9)	0,4333 (11)	0,17712 (11)	0,1882 (11)	**10,6**

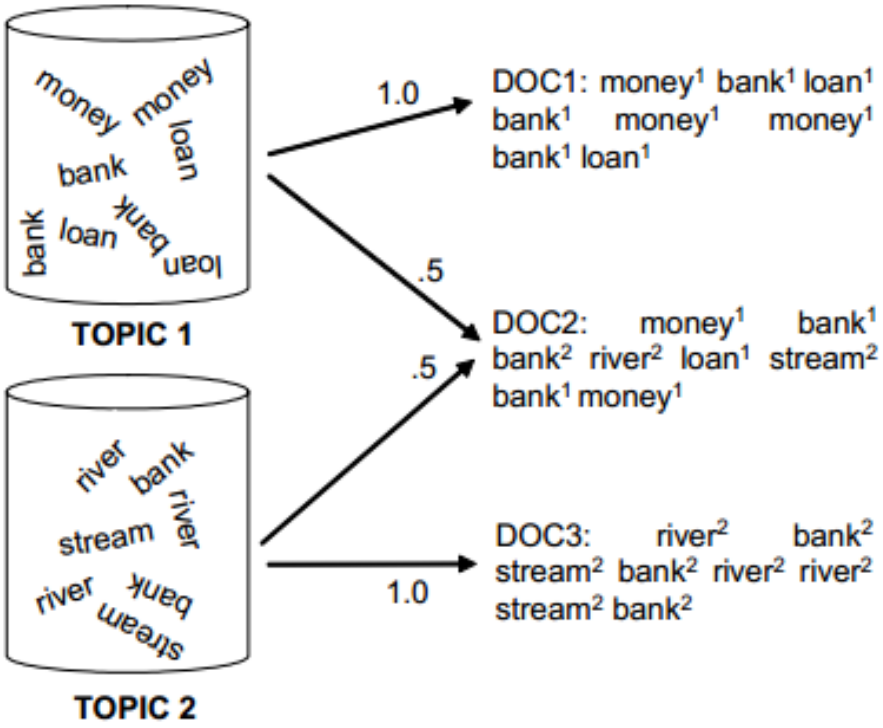
ЗАКЛЮЧЕНИЕ

- Предложена и реализована модель автоматического построения тематического представления новостного кластера, основанная на совокупности нескольких факторов
- Новым при построении тематических является разделяющий фактор: частая встречаемость выражений в одних и тех же предложениях текста
- Предложены методы интеграции тематического представления в алгоритмы автоматического аннотирования новостного кластера
- Показано улучшение качества работы алгоритмов аннотирования на основе тематического представления

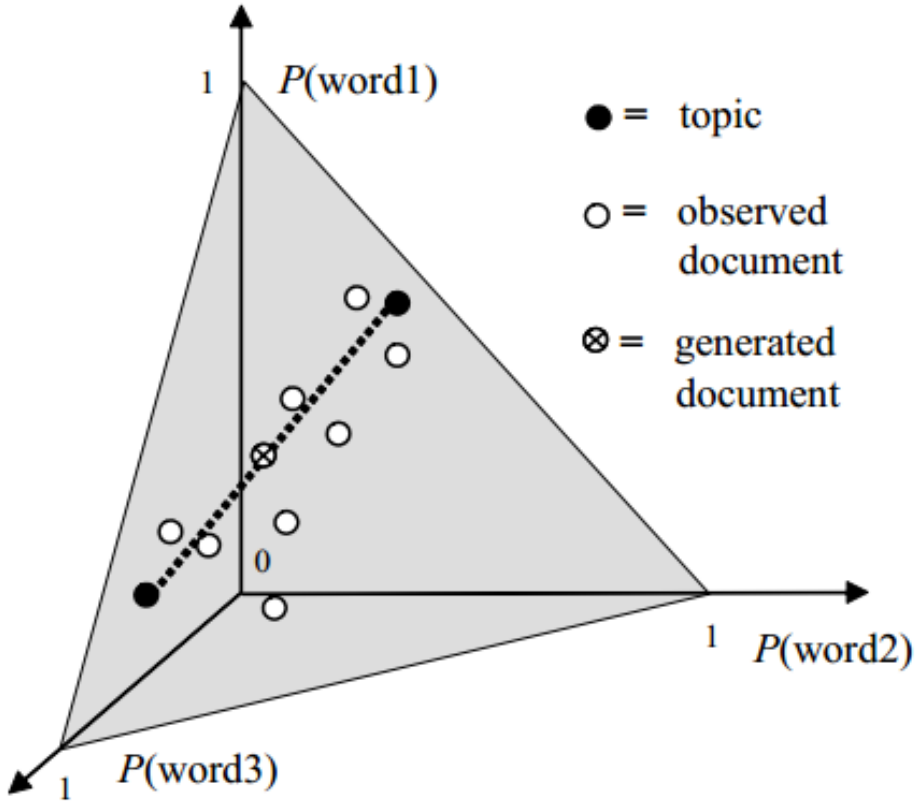
ДОПОЛНИТЕЛЬНЫЕ СЛАЙДЫ

Latent Dirichlet Allocation (LDA)

University of California, USA, 2003
David M. Blei, Andrew Y. Ng, Michael I. Jordan



Пример автоматического семплинга



Геометрическая интерпретация LDA

РАЗЛИЧИЯ С МОДЕЛЬЮ LDA

LDA	НАША МОДЕЛЬ
1. Совместная встречаемость => описание **ситуации, сферы деятельности**	1. Выделение различного именования **основных участников**
2. Результат разбиения – непрерывный (ненулевые вероятности)	2. Дискретное (строгое) разбиение
3. Не использует никаких лингвистических знаний	3. Ориентирована на аккумуляцию всех типов информации
4. Нет поддержки многословных выражений	4. Изначальное построение модели с учетом многословных выражений