Использование тематических моделей в извлечении однословных терминов

Нокель Михаил Алексеевич
Лукашевич Наталья Валентиновна

Московский Государственный Университет им. М.В. Ломоносова

План доклада

- Постановка задачи
- Коллекции текстов для экспериментов
- Статистические тематические модели
  - Основанные на методах кластеризации текстов
  - Вероятностные тематические модели
- Эксперименты
  - Выбор лучшей тематической модели
  - Сравнение тематических признаков с традиционными
- Заключение

Определение термина

Определение

Термин *– слово (или сочетание слов), являющееся точным обозначением определённого понятия какой-либо специальной области науки, техники, искусства, общественной жизни и т.п.*

Примеры терминов (из банковской области):

- *Банк*
- *Ипотечный кредит*
- *Кредит*

Определение

Извлечение терминов *– задача в автоматической обработке текстов, заключающаяся в извлечении терминов из текстов некоторой конкретной предметной области*

Автоматическое извлечение терминов

- *Приложения*:
  - Разработка тезаурусов, рубрикаторов
  - Использование в приложениях информационного поиска
  - Машинный перевод
  - Синтаксический анализ
- Эксперты используют множество принципов для определения терминов $\rightarrow$ необходимо использовать множество различных признаков для автоматического извлечения терминов
- *Текущий тренд исследований* – применение методов машинного обучения для комбинирования признаков:
  - (Pecina and Schlesinger, 2006)
  - (Foo and Merkel, 2010)
  - (Loukachevitch, 2012)

Извлечение однословных терминов

- Большинство работ посвящено извлечению многословных терминов
  - Более 85% терминов – многословные
- Мы рассматриваем *однословные* термины
  - Данная задача более трудоёмка
    - Нет внутренней структуры термина
    - Широко известные ассоциативные меры (Mutual Information, t-score и др.) не применимы
  - Одной статистики недостаточно для распознавания таких терминов в текстах
  - Необходима информация о контексте употребления

Традиционные признаки

- **Основанные на частотности**
  - *Идея*: термины встречаются чаще остальных слов
  - *TF*, *DF*, *TFIDF*, *TFRIDF* и *Domain Consensus*
- **Использующие контрастную коллекцию**
  - *Идея*: частотности терминов в целевой и контрастной коллекциях сильно различаются
  - *Weirdness*, *TFIDF*, *KFIDF*, *Loglikelihood* и др.
- **Контекстные**
  - Соединяют информацию о частотности слов с данными о контексте их употребления
  - *C-Value*, *NC-Value*, *Sum3*, *Insideness* и др.
- **Прочие**
  - Номер позиции первого вхождения в документы
  - Типы слов-кандидатов (сущ. или прил.)
  - Сущ. в именительном падеже (“подлежащие”) и др.

Основная задача

Предположение

Большинство терминов относятся к той или иной подтеме предметной области $\rightarrow$ выделение таких подтем способно улучшить качество автоматического извлечения терминов

- Необходимо исследовать возможность использования тематической информации для повышения качества извлечения однословных терминов независимо от языка и от предметной области
  - Исследовать статистические тематические модели с точки зрения задачи извлечения однословных терминов для выбора лучшей
  - Сравнить признаки, посчитанные для лучшей тематической модели, с остальными признаками для определения вклада тематической информации

Процесс извлечения терминов

Коллекции текстов для экспериментов

- **Коллекции текстов:**
  - Русскоязычные тексты из электронных банковских журналов: *Аудитор*, *Банки и Технологии*, *РБК* и др.
    - 10422 документа ($\approx$ 15.5 млн слов)
  - Английская часть корпуса *Europarl*:
    - 9673 документа ($\approx$ 54 млн слов)
- **“Золотые стандарты”:**
  - Для русского языка – банковский тезаурус, разработанный вручную для *ЦБ РФ*:
    - $\approx$ 15000 терминов
  - Для английского языка – официальный тезаурус Евросоюза *Eurovoc*:
    - 15161 термин
- **Подтверждение терминов**: слово-кандидат считается термином, если оно есть в тезаурусе

Статистические тематические модели

Определение

Статистическая тематическая модель *– модель, на основе статистических методов определяющая, к каким подтемам относится каждый документ и какие слова образуют каждую подтему*

Определение

Подтема *– список часто встречающихся рядом друг с другом слов, упорядоченный по убыванию степени принадлежности ему*

Подтема 1	Подтема 2	Подтема 3
Банкнота	*Обучение*	*Германия*
Офшорный	*Студент*	*Франция*
Счетчик	*Учебный*	*Евро*
Купюра	*Вуз*	*Европейский*

Виды статистических тематических моделей

- Тематические модели на методах кластеризации текстов
- Вероятностные тематические модели

Тематические модели на методах кластеризации

- Основываются на методах жесткой кластеризации
  - Каждый документ – разреженный вектор в пространстве слов большой размерности
- В конце кластеризации каждый кластер – один большой документ для вычисления вероятностей слов:

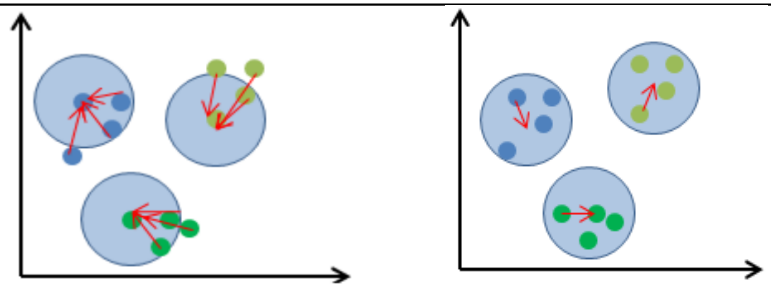
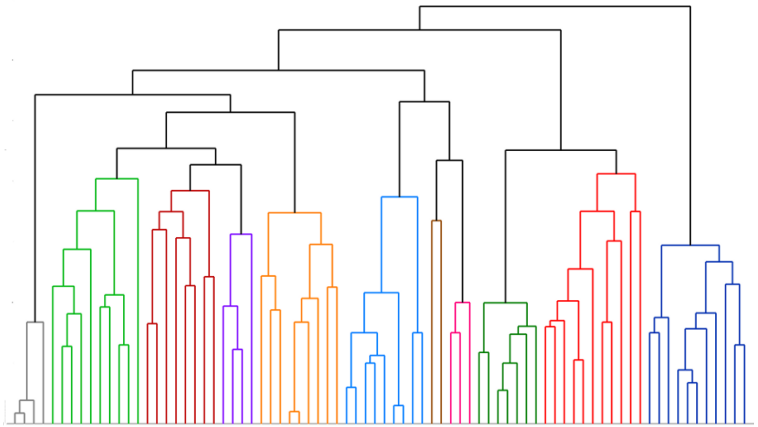
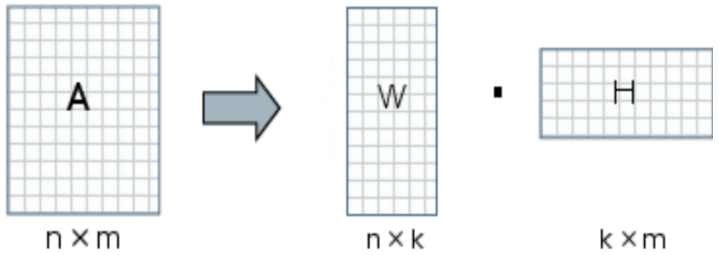
$$P(w|t) = \frac{TF(w|t)}{\sum_{w} TF(w|t)}$$

- Общие шаги в процессе кластеризации:
  - Предобработка документов (фильтрация слов)
  - Преобразование документов в вектор слов
  - Расчет расстояния между документами
    - Схема взвешивания слов:

$$TFIDF(w|d) = TF(w|d) \times \max\left(0, \log \frac{N - DF(w)}{DF(w)}\right)$$

  - Кластеризация документов

Тематические модели на методах кластеризации

K-Средних • K-Средних • Сферический K-Средних	
Агломеративная кластеризация • Single-link • Complete-link • Average-link	
Неотрицательная матричная факторизация (NMF) • NMF Euclidean • NMF KL-Divergence	

Вероятностные тематические модели

- Каждый документ – смесь подтем
- Каждая подтема – вероятностное распределение над словами
- Вероятностная модель порождения документа d:

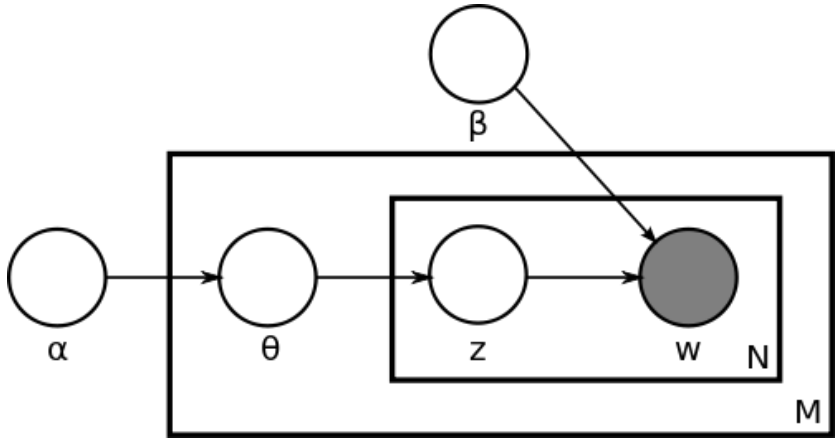
$$P(w|d) = \sum_t P(w|t)P(t|d)$$

- Процесс порождения слов:

Для всех документов $d \in D$:
Для всех слов $w \in d$:
Выбрать тему t из $p(t|d)$
Выбрать слово w из $p(w|t)$

Вероятностные тематические модели

- Метод вероятностного латентного семантического моделирования (*PLSA*)
  - (Ding, Li, Peng, 2008) показали, что NMF, минимизирующий расстояние Кульбака-Лейблера, эквивалентен PLSA
- Латентное размещение Дирихле (*LDA*)



  - Приближённые способы настройки параметров:
    - *LDA VB* – оригинальная реализация Blei
    - *LDA Gibbs* – реализация LDAGibbs++

Признаки, использующие тематическую информацию

Предположение

В начале списков, образующих подтемы, с большой вероятностью находятся термины

Признак	**Формула**
Частотность (TF)	$\sum_t P(w\|t)$
TFIDF	$TF(w) \times \log \frac{K}{DF(w)}$
Domain Consensus (DC)	$-\sum_t P(w\|t) \times \log P(w\|t)$
Maximum TF	$\max_t P(w\|t)$
Term Score (TS)	$\sum_t TS(w\|t)$, $TS(w\|t) = P(w\|t) \times \log \frac{P(w\|t)}{\left(\prod_t P(w\|t)\right)^{\frac{1}{K}}}$
TS-IDF	$TS(w) \times \log \frac{K}{DF(w)}$
Maximum TS	$\max_t TS(w\|t)$

Baseline и метрика оценки качества

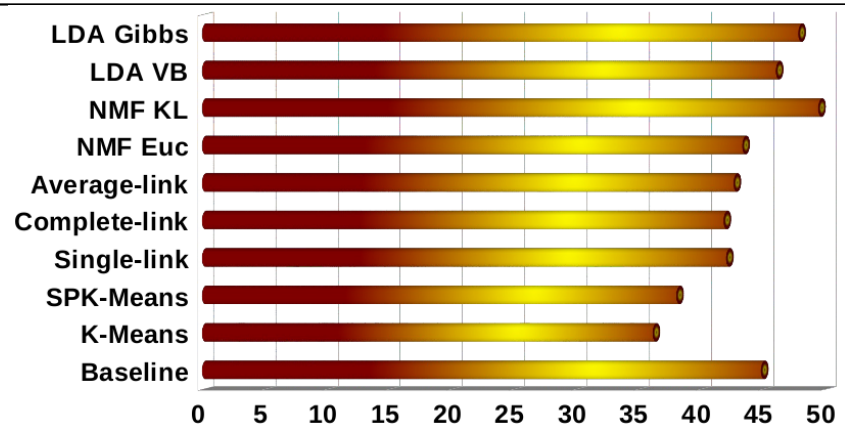
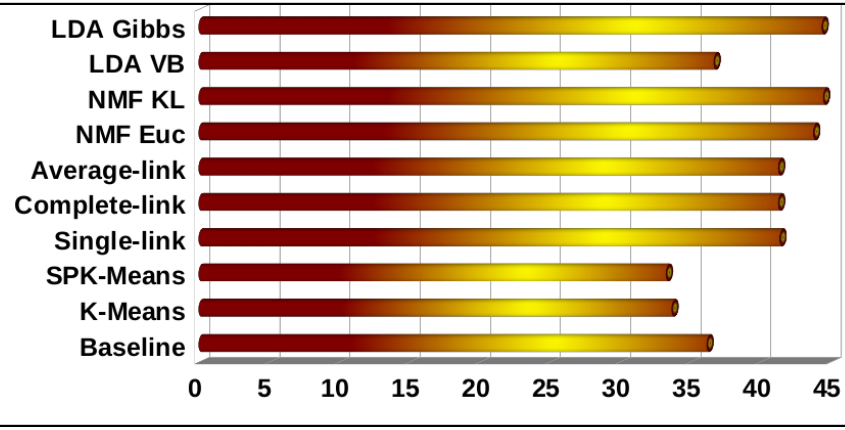
- Baseline – “тематическая модель”
  - Не выделяет никаких подтем
  - Каждый документ – отдельная подтема
- Метрика оценки качества – Средняя Точность (AvP):

$$AvP(n) = \frac{\sum_{k=1}^{n} Precision(k)}{number\ of\ terms}$$

- Разное число подтем (50, 100 и 150) на качество извлечения терминов никак не повлияло $\rightarrow$ число подтем равно 100
- Все признаки рассчитывались для 5000 самых частотных слов

Выбор лучшей тематической модели

- Комбинирование тематических признаков для каждой модели логистической регрессией (библиотека Weka)
- Четырёхкратная кросс-проверка

Русская банковская коллекция	
Английская коллекция Europarl	

- **Вывод**: лучшая тематическая модель – **NMF KL**

Сравнение тематических признаков с традиционными

- **Традиционные признаки**
  - **Основанные на частотности**
    - *TF*, *DF*, *TFIDF*, *TFRIDF* и *Domain Consensus*
  - **Использующие контрастную коллекцию**
    - Корпус русского языка
    - Британский национальный корпус
    - *Weirdness*, *TFIDF*, *KFIDF*, *Loglikelihood* и др.
  - **Контекстные**
    - *C-Value*, *NC-Value*, *Sum3*, *Insideness* и др.
  - **Прочие**
    - Номер позиции первого вхождения в документы
    - Сущ. в именительном падеже (“подлежащие”) и др.
- **Тематические признаки** (для модели **NMF KL**)
  - *TF*, *TFIDF*, *Domain Consensus*, *Maximum TF*, *TS*, *TS-IDF*, *Maximum TS*

Лучшие признаки

- Русская банковская коллекция:

Группа признаков	**Лучший признак**	**AvP**
Основанные на частотности	*TFRIDF*	41.1
Использующие контрастную коллекцию	*LogLikelihood*	36.9
Контекстные	*Sum3*	37.4
Тематические	**Term Score**	**48.9**

- Английская коллекция Europarl:

Группа признаков	**Лучший признак**	**AvP**
Основанные на частотности	*TFRIDF* для подлежащих	38.5
Использующие контрастную коллекцию	*TFIDF* для подлежащих	34.2
Контекстные	*C-Value*	31.3
Тематические	**Term Score**	**44.5**

- **Вывод**: независимо от языка и предметной области лучшие признаки – тематические (+19% и +15% AvP по сравнению с остальными)

Определение вклада тематических признаков

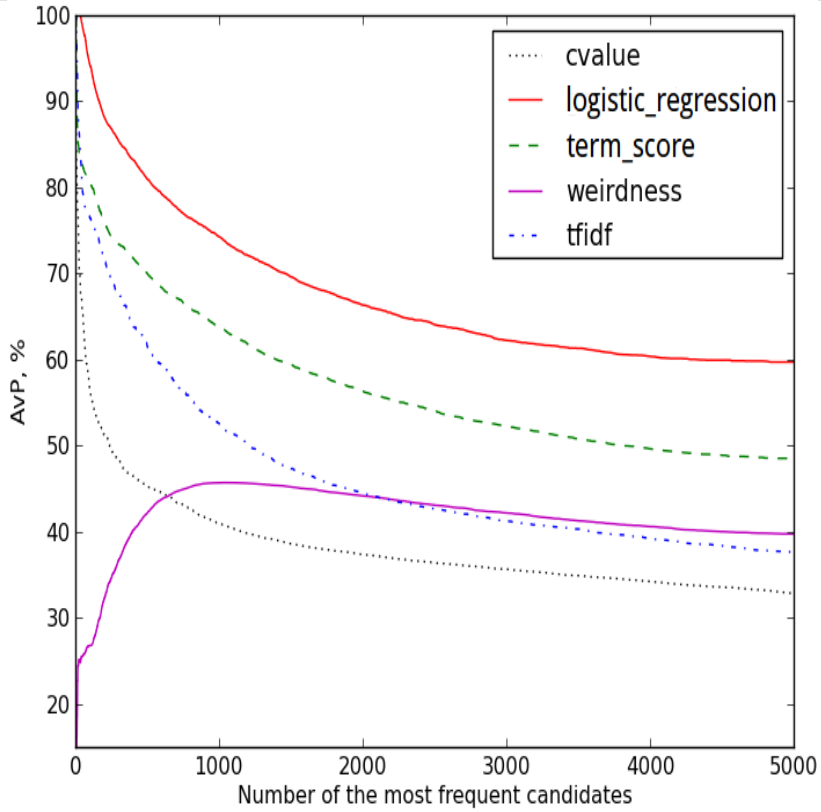
- Модель с тематическими признаками (7 baseline признаков + 7 признаков, посчитанных для NMF KL) vs. модель без них
- Комбинирование признаков – методом логистической регрессии (библиотека Weka)
- **Результаты**:

Коллекция	**Средняя точность**	
	Без тематических признаков	С тематическими признаками
Русская	54.6	**56.3**
Английская	50.4	**51.4**

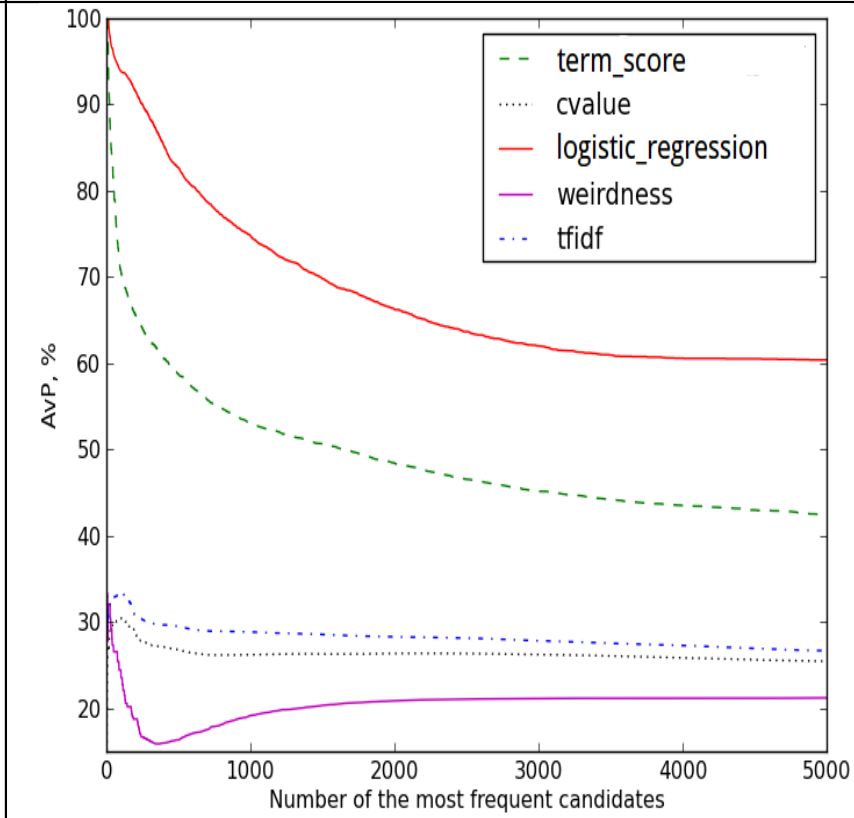
- **Вывод**: тематические модели вносят дополнительную информацию в процесс автоматического извлечения терминов

Графики средней точности

Русская банковская коллекция



Английская коллекция Europarl



Примеры извлечённых терминов

- Модель с тематическими признаками
- Подтверждённые термины выделены ***цветом***

№	Русский корпус	Английский корпус
1	***Банковский***	Member
2	***Банк***	Minute
3	Год	***Amendment***
4	***РФ***	***Document***
5	***Кредитный***	EU
6	***Налоговый***	President
7	***Кредит***	***People***
8	***Пенсионный***	***Directive***
9	Средство	Year
10	Клиент	Question

Заключение

- Предложено использовать тематические модели для извлечения терминов
- На основе тематических моделей предложено несколько модификаций известных признаков для ранжирования слов-кандидатов по степени их терминологичности
- Экспериментально показано, что использование тематической информации способно улучшить качество автоматического извлечения однословных терминов независимо от предметной области и языка

Спасибо за внимание!

Вопросы?

Традиционные признаки

- Основанные на частотности:

$$TFIDF(w) = TF(w) \times \log \frac{|D|}{DF(w)}$$

$$TFRIDF(w) = TF(w) \times \left(\log \frac{|D|}{DF(w)} - \left(-\log \left(1 - e^{-\frac{TF(w)}{|D|}} \right) \right) \right)$$

$$DC(w) = -\sum_{d \in D} (TF(w|d) \times \log TF(w|d))$$

- Использующие контрастную коллекцию:

$$Weirdness(w) = \left(\frac{TF_t(w)}{|W_t|} \right) / \left(\frac{TF_r(w)}{|W_r|} \right)$$

$$KFIDF(w) = DF_t(w) \times \left(\frac{2}{|D|_w} + 1 \right)$$

где $|D|_w = 1$, если слово содержится в контрастной коллекции и $|D|_w = 2$ иначе

Традиционные признаки

- Контекстные:

$$C - Value(w) = TF(w) - \frac{\sum_{p \in P_w} TF(p)}{|P_w|}$$

где P_w – множество объемлющих фраз, содержащих слово w

$$SumN(w) = \frac{\sum\limits_{p \in P_w^N} TF(p)}{N}$$

где P_w^N – множество N самых частотных объемлющих фраз, содержащих слово w

$$Insideness(w) = \frac{F_{max}}{TF_t(w)}$$

где F_{max} – максимальная частота объемлющей фразы, содержащей слово w