Использование тематических моделей в извлечении однословных терминов

© М. А. Нокель
МГУ им. М. В. Ломоносова, Москва
mnokel@gmail.com

© Н. В. Лукашевич
НИВЦ МГУ им. М. В. Ломоносова, Москва
louk_nat@mail.ru

Аннотация

В статье представлены результаты экспериментов по применению тематических моделей к задаче извлечения однословных терминов. В качестве текстовых коллекций была взята подборка статей из электронных банковских журналов на русском языке и англоязычная часть корпуса параллельных текстов Europarl. Эксперименты показывают, что использование тематической информации значительно улучшает качество извлечения однословных терминов независимо от предметной области и используемого языка.

Ключевые слова

Тематические модели, Кластеризация, Извлечение однословных терминов

1 Введение

Извлечение терминов из текстов определённой предметной области играет значительную роль во многих прикладных задачах, в первую очередь – в разработке и пополнении различных терминологических ресурсов, таких как тезаурусы и онтологии [35]. Поскольку разработка таких ресурсов вручную достаточно трудоёмка, за последние годы было проведено большое количество исследований по автоматизации данного процесса.

Большинство современных методов извлечения терминов основываются на использовании различных статистических и лингвистических признаков слов. Основная цель при этом заключается в получении упорядоченного списка кандидатов в термины, в начале которого находится как можно больше слов, с наибольшей вероятностью являющихся терминами. В некоторых работах было экспериментально установлено, что использование машинного обучения для комбинирования признаков значительно улучшает результаты извлечения терминов по сравнению с методами, основанными только на одном каком-то признаке, поскольку те или иные признаки только частично отражают особенности поведения терминов в текстах [17].

На текущий момент традиционно используемые для извлечения терминов статистические признаки никак не отражают тот факт, что большинство терминов относятся к той или иной подтеме предметной области. Поэтому нами было сделано предположение, что выделение таких подтем в коллекции текстов способно улучшить качество автоматического извлечения терминов.Для проверки этого предположения в статье будут рассмотрены различные методы выделения подтем в коллекции текстов, которые часто в литературе называются статистическими тематическими моделями [4].

Некоторые виды статистических тематических моделей могут основываться на традиционных методах автоматической кластеризации текстов [12]. В последнее время предложены вероятностные механизмы выделения подтем в текстовых коллекциях такие, как методы, основанные на скрытом распределении Дирихле (Latent Dirichlet allocation [4]), которые собственно и были названы тематическими моделями и в настоящее время интенсивно исследуются в рамках различных приложениях автоматической обработки текстов ([12], [29], [3]).

Основная задача данной статьи заключается в исследовании возможности использования тематической информации для повышения качества извлечения однословных терминов. Для этой цели вначале в текстовой коллекции выделяются подтемы, затем к ним применяются некоторые модификации хорошо известных признаков, которые впоследствии используются вместе с другими статистическими и лингвистическими признаками.

Для того чтобы результаты, представленные в статье, не зависели ни от предметной области, ни от языка, были взяты две предметные области и соответствующие текстовые коллекции: банковская предметная область и тексты банковской тематики на русском языке и широкая предметная область современной общественной жизни Европы и речи с заседаний Европарламента на английском языке. При этом эксперименты будут строиться следующим образом:

1. Вначале статистические тематические модели будут исследованы с точки зрения задачи

извлечения однословных терминов с целью выбора наилучшей;

2. Затем будет осуществлено сравнение признаков, посчитанных для лучшей тематической модели, с остальными признаками с целью определения вклада, который даёт использование тематической модели в рассматриваемой задаче.

2 Близкие работы

За последние годы было предложено много различных статистических и лингвистических признаков слов, используемых для извлечения однословных терминов из коллекции текстов определённой предметной области ([6], [1], [20], [10] и др.).

Все предложенные признаки можно разделить на следующие группы:

1. *Признаки, основанные на частотности слов-кандидатов.* К этой группе относится, например, признак *TFRIDF*, предложенный в работе [6] и использующий модель Пуассона для предсказания терминологичности слов;

2. *Признаки, использующие контрастную коллекцию*, т.е. коллекцию более общей тематики. Одним из наиболее характерных представителей данной группы является широко используемый на практике признак *Относительная частотность* [1], основанный на сравнении относительных частотностей слов в рассматриваемой и в контрастной текстовой коллекциях;

3. *Контекстные признаки*, соединяющие в себе информацию о частотности слов-кандидатов с данными о контексте их употребления. Наиболее известными представителями этой группы являются признаки *C-Value* [20] и *NC-Value* [10], учитывающие частоту встречаемости объемлющего словосочетания для кандидата в термины.

Однако ни один из предложенных признаков не является определяющим [25], и фактически из текстов извлекается довольно большой список слов-кандидатов, которые затем должны быть проанализированы и подтверждены экспертом по предметной области. Важно поэтому дополнять список используемых признаков, что позволит получать в начале списка как можно больше слов, с наибольшей вероятностью являющихся терминами.

3 Статистические тематические модели

Новые признаки слов-кандидатов, которые вводятся в данной статье, используют информацию, получаемую статистическими тематическими моделями в исследуемых текстовых коллекциях.

Статистическая тематическая модель (далее – тематическая модель) коллекции текстовых документов на основе статистических методов определяет, к каким подтемам относится каждый документ и какие слова образуют каждую подтему, представляющую собой список часто встречающихся рядом друг с другом слов, упорядоченный по убыванию степени принадлежности ему [34]. Так, в таблице 1 представлены первые десять слов, наиболее полно характеризующие три случайно выбранных подтемы, выделенных из русскоязычных текстов банковской тематики рассматриваемой коллекции.

Подтема 1	**Подтема 2**	**Подтема 3**
Банкнота	*Обучение*	*Германия*
Офшорный	*Студент*	*Франция*
Счетчик	*Учебный*	*Евро*
Купюра	*Вуз*	*Европейский*
Подделка	*Семинар*	*Польша*
Обращение	*Образование*	*Европа*
Номинал	*Знание*	*Чехия*
Монета	*Специалист*	*Италия*
Подлинность	*Слушатель*	*Немецкий*
Поддельный	*Учитель*	*Французский*

Таблица 1: Примеры подтем

В тематических моделях, как правило, используется модель мешка слов, в которой каждый документ рассматривается как набор встречающихся в нём слов. При этом перед выделением подтем текстовая коллекция обычно подвергается предобработке, выделяющей только значимые слова в каждом документе. В частности, в данном исследовании для русского языка были отобраны только существительные и прилагательные, а для английского – только существительные, поскольку они покрывают большую часть терминов.

На сегодняшний день разработано достаточно много различных тематических моделей. Для выбора моделей для исследования были проанализированы предыдущие работы, в которых осуществляется сравнение моделей с точки зрения различных практических приложений. Так, в работе [29] утверждается, что каждая тематическая модель имеет свои сильные и слабые стороны. Сравнивая между собой методы NMF (неотрицательной матричной факторизации) и LDA (латентного размещения Дирихле), авторы приходят к выводу, что оба этих алгоритма дают похожее ка-

чество, хотя NMF и выдаёт немного больше бессвязных подтем. В работе [12] утверждается, что традиционные тематические модели показывают приемлемое качество выделения подтем, но имеют множество ограничений. В частности они предполагают, что каждый документ имеет только одну тематику. В действительности же документы представляют собой, как правило, смесь подтем. Кроме того, авторы отмечают, что параметры традиционных моделей достаточно сложно настраивать. В то же время в работе подчёркивается, что более сложные модели (такие как LDA) необязательно дадут лучшие результаты.

Поскольку, как следует из упомянутых выше работ, среди тематических моделей нет явного лидера и непонятно, какое качество они покажут в рассматриваемой задаче извлечения однословных терминов, было решено выбрать несколько наиболее характерных представителей, которых условно можно отнести либо к вероятностным, либо к методам кластеризации текстов, рассматриваемых с точки зрения тематических моделей. Каждая из выбранных моделей будет рассмотрена в следующих подразделах.

3.1 Тематические модели, основанные на методах кластеризации текстов

Традиционные тематические модели, как правило, основываются на методах жёсткой кластеризации, рассматривающих каждый документ как разреженный вектор в пространстве слов большой размерности [28]. После окончания работы алгоритма кластеризации каждый получившийся кластер рассматривается как один большой документ для вычисления вероятностей входящих в него слов по следующей формуле:

$$P(w|t) = \frac{TF(w|t)}{\sum\limits_{w} TF(w|t)} \tag{1}$$

где $TF(w|t)$ – частотность слова w в кластере t.

В процессе кластеризации текстовых документов можно выделить следующие общие шаги:

1. Предобработка документов (фильтрация слов);
2. Преобразование документа во внутреннее представление (в вектор слов);
3. Расчёт расстояния между документами на основе внутреннего представления;
4. Кластеризация документов на основе рассчитанного расстояния с помощью одного из алгоритмов.

Для численной оценки расстояния между документами необходим способ определения значимости каждого слова в обособлении одного документа относительно другого. Для этого были предложены различные схемы взвешивания отдельных слов, наиболее распространённой из которых является схема TFIDF [19], которая также была включена в данное исследование. В ней каждому слову в документе ставится в соответствие величина, вычисляемая по следующей формуле:

$$TFIDF(w|d) = TF(w|d) * \max\left(0, \log\frac{N - DF(w)}{DF(w)}\right) \tag{2}$$

где N – общее число документов в коллекции, $DF(w)$ – число документов в коллекции, в которых встречается слово w.

В следующих разделах будут описаны выбранные нами методы построения традиционных тематических моделей.

3.1.1 K-Средних и Сферический K-Средних

Алгоритм K-Средних [18] начинает свою работу со случайной инициализации центров масс каждого кластера. Далее он итеративно повторяет следующие шаги:

1. Все документы разбиваются на кластеры в соответствии с тем, какой из центров масс оказался ближе по выбранной метрике;
2. Для каждого полученного кластера пересчитывается центр масс.

В качестве метрики близости между двумя документами исследовались следующие:

- Евклидово расстояние (*K-Means*) [18]:

$$sim(A, B) = \sqrt{\sum_{i}(A_i - B_i)^2} \tag{3}$$

- Косинусная мера близости *(сферический k-средних – SPK-Means)*. При этом все векторы, представляющие документы, нормализуются к единичной гиперсфере [33]:

$$sim(A, B) = \frac{\sum\limits_{i}(A_i \times B_i)}{\sqrt{\sum\limits_{i} A_i} \times \sqrt{\sum\limits_{i} B_i}} \tag{4}$$

3.1.2 Иерархическая агломеративная кластеризация

Алгоритм иерархической агломеративной кластеризации [14] изначально рассматривает каждый документ как отдельный кластер. Затем он итеративно повторяет следующие шаги:

1. Находятся и объединяются в новый кластер два наиболее близких кластера;
2. Пересчитываются расстояния между новым кластером и всеми остальными.

Процесс повторяется до тех пор, пока не останется заданное число кластеров.

В качестве способов определения наиболее близких кластеров исследовались следующие наиболее распространённые [14]:

- *Complete-link* ("полное связывание"). Наиболее близкие кластеры – это кластеры с наименьшим максимальным парным расстоянием между документами;
- *Single-link* ("одиночное связывание"). Наиболее близкие кластеры – это кластеры с наименьшим минимальным парным расстоянием между документами;
- *Average-link* ("среднее связывание"). Это компромисс между двумя предыдущими способами. Наиболее близкие кластеры – это кластеры с наименьшим средним парным расстоянием между документами.

3.1.3 Неотрицательная матричная факторизация (NMF)

Алгоритм NMF, изначально разработанный для уменьшения размерности, зарекомендовал себя для решения задач кластеризации [32]. Данный алгоритм осуществляет нечёткую кластеризацию, которая относит один и тот же документ к разным кластерам с разными вероятностями.

Принимая на входе неотрицательную разреженную матрицу V, которая получается записыванием векторов, представляющих документы, по столбцам, алгоритм ищет такие матрицы W и H меньшей размерности, что $V \approx WH$ по некоторой метрике. В качестве такой метрики исследовались следующие [16]:

- Евклидово расстояние (*NMF Euc*):

$$||A - B||^2 = \sum_{i,j}(A_{ij} - B_{ij})^2 \qquad (5)$$

- Расстояние Кульбака-Лейблера для неотрицательных матриц (*NMF KL*):

$$D(A||B) = \sum_{i,j}(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}) \quad (6)$$

В результате работы алгоритма в матрице W получается распределение слов по кластерам, а в матрице H – распределение документов по кластерам. Нормируя соответствующие величины для каждого слова/документа, можно получить вероятности принадлежности этого слова/документа кластеру.

3.2 Вероятностные тематические модели

Вероятностные тематические модели представляют каждый документ в виде смеси подтем, в которой каждая подтема представляет собой некоторое вероятностное распределение над словами. Вероятностные модели порождают слова по следующему правилу:

$$P(w|d) = \sum_t P(w|t)P(t|d) \qquad (7)$$

где $P(t|d)$ и $P(w|t)$ – распределения подтем по документам и слов по подтемам, а $P(w|d)$ – наблюдаемое распределение слов по документам.

Порождение слов происходит следующим образом. Для каждого документа d и для каждого слова $w \in d$ выбирается тема t из распределения $P(t|d)$, и затем генерируется слово w из распределения $P(w|t)$.

Самыми известными представителями данной категории являются метод вероятностного латентного семантического индексирования (PLSI) и латентное размещение Дирихле (LDA).

3.2.1 PLSI

Метод PLSI, также известный как PLSA, был предложен в работе [13]. Данный метод моделирует матрицу V, в которой V_{ij} обозначает число вхождений слова w_i в документ d_j, получающуюся из модели с k подтемами:

$$P(w_i, d_j) = \sum_{t=1}^{k} P(t)P(d_j|t)P(w_i|t) \qquad (8)$$

Параметры модели настраиваются с помощью максимизации правдоподобия наблюдаемых данных из матрицы M, т.е. максимизируя следующий функционал:

$$\sum_{i,j} TF(w_i|d_j) \log P(w_i, d_j) \to \max \qquad (9)$$

Поскольку в статье [7] теоретически обосновано, что алгоритм NMF, минимизирующий расстояние Кульбака-Лейблера и рассмотренный в прошлом разделе, эквивалентен алгоритму PLSA, в данном исследовании метод PLSA не рассматривается отдельно.

3.2.2 LDA

Метод латентного размещения Дирихле был предложен в работе [4]. LDA расширяет модель PLSI, добавляя туда априорное распределение параметров модели ($P(w|t)$ и $P(t|d)$), считая их распределёнными по закону Дирихле.

Для настройки параметров модели необходим Байесовский вывод. Однако, поскольку он алгоритмически неразрешим [4], исследовались следующие два применяемых на практике приближённых способа Байесовского вывода:

- *LDA VB* – вариационный Байесовский вывод, описанный в статье [4];
- *LDA Gibbs* – метод Монте-Карло с марковскими цепями, использующий сэмплирование Гиббса [27].

3.3 Базовая тематическая модель

В качестве baseline была взята "тематическую" модель, которая не выделяет никаких подтем, а просто рассматривает каждый документ как отдельно взятую подтему. Данная модель будет использоваться нами в экспериментах для сравнения с другими методами.

4 Коллекции текстов для экспериментов

Во всех экспериментах, описываемых в данной статье, слова-кандидаты извлекались из двух различных коллекций:

- Коллекция банковских русскоязычных текстов (10422 документа, примерно 15.5 млн слов), взятых из различных электронных банковских журналов: Аудитор, Банки и Технологии, РБК и др.;
- Английская часть корпуса параллельных текстов Europarl [8] из заседаний Европарламента (9673 документа, примерно 54 млн слов).

Для подтверждения терминологичности слов-кандидатов использовались следующие "золотые стандарты":

- Для русского языка – тезаурус, разработанный вручную для Центрального Банка Российской Федерации и включающий в себя порядка 15000 терминов, относящихся к сфере банковской активности, денежной политики и макроэкономики;
- Для английского языка – официальный многопрофильный тезаурус Европейского Союза Eurovoc [9], предназначенный для ручного индексирования заседаний Европарламента. Его английская версия включает в себя 15161 термин.

При этом слово-кандидат считается термином, если оно содержится в тезаурусе.

Все признаки слов-кандидатов рассчитывались для 5000 самых частотных слов. В качестве метрики оценки качества была выбрана Средняя Точность (AvP) [19], определяемая для множества D всех слов-кандидатов и его подмножества $D_q \subseteq D$, представляющего действительно термины (т.е. подтверждённые тезаурусом):

$$AvP(n) = \frac{1}{|D_q|} \sum_{1 \le k \le |D_q|} \left(r_k \times \left(\frac{1}{k} \sum_{1 \le i \le k} r_i \right) \right) \quad (10)$$

где $r_i = 1$, если i-е слово-кандидат $\in D_q$, и $r_i = 0$ иначе. Данная формула отражает тот факт, что чем больше терминов сосредоточено в вершине итогового списка слов-кандидатов, тем выше мера средней точности.

Эксперименты проводились с разным числом выделяемых подтем: 50, 100 и 150 соответственно. Визуально результаты получались разными, но на качестве извлечения терминов это никак не отразилось. Поэтому все дальнейшие эксперименты проводилось с числом подтем, равным 100.

5 Выбор лучшей тематической модели

Как уже было сказано выше, вначале будут представлены результаты экспериментов по определению наилучшей тематической модели. Для этого будут предложены и посчитаны для каждой из рассмотренных выше тематических моделей некоторые модификации известных признаков слов.

5.1 Признаки, использующие тематическую информацию

Основной идеей всех признаков, использующих полученную с помощью какой-либо тематической модели информацию, является тот факт, что в начале списков, образующих подтемы, с большой вероятностью находятся термины. Для экспериментов мы предложили некоторые модификации известных признаков (см. таблицу 2). В таблице 2 используются следующие обозначения:

- $TF(w)$ – частотность слова w
- $DF(w)$ – документная частотность слова w
- $P(w|t)$ – условная вероятность принадлежности слова w подтеме t
- k – число топиков

5.2 Результаты экспериментов

В таблицах 3 и 4 представлены результаты экспериментов для исследуемых русского и английского корпуса соответственно.

Как видно из приведённых выше таблиц, лучшее качество независимо от языка и предметной области даёт тематическая модель **NMF**, минимизирующая расстояние Кульбака-Лейблера. Так, лучшим признаком для обоих языков является *Term Score* с 16% (соответственно 21%) прироста

Признак	Формула
Частотность (TF)	$\sum_t P(w\|t)$
TFIDF	$TF(w) \times \log \frac{k}{DF(w)}$
Domain Consensus (DC) [22]	$-\sum_t (P(w\|t) \times \log P(w\|t))$
Maximum TF	$\max_t P(w\|t)$
Term Score (TS) [3]	$\sum_t TS(w\|t)$ $TS(w\|t) = P(w\|t) \log \frac{P(w\|t)}{(\prod_t P(w\|t))^{\frac{1}{k}}}$
TS-IDF	$TS(w) \times \log \frac{k}{DF(w)}$
Maximum TS (MTS)	$\max_t TS(w\|t)$

Таблица 2: Признаки, использующие тематическую информацию

Модель	TF	TFIDF	DC	MTF	TS	TSIDF	MTS
K-Means	33.3	25.5	32.7	34.4	35.7	28.7	34.3
SPK-Means	35.5	27.2	35	33.9	36.3	30.1	33.6
Single-link	34.8	39.9	33.6	38.9	38.4	40.5	39
Comp-link	35.6	41	34.5	39.2	38.4	41	39.5
Average-link	35.8	40.7	34.5	39.5	39	40.9	39.6
NMF Euc	40.8	42.5	40.3	40.8	42	43.1	41.9
NMF KL	42.3	40.3	37.5	47.1	**48.9**	42.9	47.9
LDA VB	35.8	42.7	32.8	42.8	42.5	45.1	46.5
LDA Gibbs	37.7	38.4	35	46.2	42.6	42.8	47.2
Baseline	34	37.6	32.8	38.5	38.1	42	38.1

Таблица 3: Средняя точность признаков на русском корпусе

Model	TF	TFIDF	DC	MTF	TS	TSIDF	MTS
K-Means	29.3	32.3	28.9	30.3	30.1	31.8	30.4
SPK-Means	28.1	29.8	27.9	28.7	28.6	29.7	28.7
Single-link	30.3	38.9	29.8	37.3	36.5	38.8	39.9
Comp-link	31.1	39.6	30.4	37.2	34.6	38.9	39
Average-link	30.5	38.9	29.9	37.1	35.4	38.3	39.3
NMF Euc	34.4	31.6	32.3	41.1	43.7	31.6	40.5
NMF KL	33.3	37.7	31.2	44.3	**44.4**	37.3	44.1
LDA VB	32.3	30.3	30.5	37.1	36.3	30.3	38.5
LDA Gibbs	35.2	41.8	33.3	42.6	37.8	43.7	43.5
Baseline	31.5	32.8	30	36	33.6	35	36.7

Таблица 4: Средняя точность признаков на английском корпусе

качества относительно лучших признаков базовой модели (*TFIDF* для русского корпуса и *Maximum Term Score* для английского корпуса).

Помимо вычисления средней точности отдельных признаков было также осуществлено их комбинирование для каждой исследуемой тематической модели в отдельности с помощью метода логистической регрессии, реализованного в библиотеке Weka [30]. При этом проводилась четырёхкратная кросс-проверка, означающая, что вся исходная выборка разбивалась случайным образом на четыре равные непересекающиеся части, и каждая часть по очереди становилась контрольной подвыборкой, а обучение проводилось по остальным трём. Результаты комбинирования признаков для русского и английского корпусов представлены в таблице 5.

Как видно из приведённых выше таблиц, тематическая модель **NMF**, минимизирующая расстояние Кульбака-Лейблера, снова даёт наилучшее качество с 10% прироста для русского и с 23% прироста для английского корпусов относительно базовой тематической модели.

Модель	Средняя точность	
	Для русского корпуса	Для английского корпуса
Baseline	44.9	36.2
K-Means	36.2	33.7
SPK-Means	38.1	33.3
Single-link	42.1	41.4
Complete-link	41.9	41.3
Average-link	42.7	41.3
NMF Euc	43.4	43.8
NMF KL	**49.5**	**44.5**
LDA VB	46.1	36.7
LDA Gibbs	47.9	44.4

Таблица 5: Средняя точность комбинирования признаков, использующих тематическую информацию

Таким образом, наилучшей тематической моделью оказалась модель **NMF**, минимизирующая расстояние Кульбака-Лейблера.

6 Сравнение с другими признаками

Для изучения вклада тематической информации в задачу автоматического извлечения однословных терминов было решено сравнить результаты предложенных признаков, использующих тематическую информацию, с остальными статистическими и лингвистическими признаками для обоих исследуемых корпусов для 5000 самых частотных слов.

В качестве признаков, не использующих тематическую информацию, были взяты характерные представители групп, описанных в разделе 2.

6.1 Признаки, основанные на частотности

Признаки из данной группы опираются на предположение о том, что термины, как правило, встречаются в коллекции гораздо чаще остальных слов. В исследование были включены следующие признаки: *Частотность*, *Документная частотность*, *TFIDF* [19], *TFRIDF* [6], *Domain Consensus* [22].

6.2 Признаки, использующие контрастную коллекцию

Для вычисления признаков этой категории помимо целевой коллекции текстов предметной области использовалась контрастная коллекция текстов более общей тематики. Для русского языка в качестве таковой была взята подборка из примерно 1 миллиона новостных текстов, а для англий-

ского – n-граммные статистики из Британского Национального Корпуса [5].

Основная идея таких признаков заключается в том, что частотности терминов в целевой и контрастной коллекциях существенно различаются. В данном исследовании рассматривались следующие признаки: *Относительная частотность* [1], *Релевантность* [26], *TFIDF* [19] с вычислением документной частотности по контрастной коллекции, *Contrastive Weight* [2], *Discriminative Weight* [31], *KF-IDF* [15], *Lexical Cohesion* [24] и *Логарифм правдоподобия* [11].

6.3 Контекстные признаки

Контекстные признаки соединяют в себе информацию о частотности слов-кандидатов с данными о контексте их употребления в коллекции. В данном исследовании рассматривались следующие признаки: *C-Value* [20], *NC-Value*, *MNC-Value* [10], *Token-LR*, *Token-FLR*, *Type-LR*, *Type-FLR* [21], *Sum3*, *Sum10*, *Sum50*, *Insideness* [17].

6.4 Прочие признаки

В качестве остальных признаков, не использующих тематическую информацию, рассматривались номер позиции первого вхождения в документы, типы слов-кандидатов (существительное или прилагательное), слова-кандидаты, начинающиеся с заглавной буквы, и существительные в именительном падеже ("подлежащие") и слова из контекстного окна с некоторыми самыми частотными предопределёнными терминами [23].

Кроме этого, также рассматривались и комбинации данных признаков с некоторыми статистическими величинами (такими, как частотность в целевом корпусе). Всего было взято 28 таких признаков.

6.5 Результаты экспериментов

Лучшие признаки каждой из упомянутых выше групп для русского и английского корпусов приведены в таблицах 6 и 7.

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

Таблица 6: Средняя точность лучших признаков для русского корпуса

Как видно из приведённых выше таблиц, независимо от языка и предметной области лучшими индивидуальными признаками оказались тематические, превзойдя остальные на 19% и 15% средней точности для русского и английского корпусов соответственно.

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**

Таблица 7: Средняя точность лучших признаков для английского корпуса

Для оценки же вклада тематических признаков в общую модель извлечения однословных терминов мы сравнили модель извлечения, учитывающую тематические признаки (7 baseline признаков и 7 признаков, посчитанных для наилучшей тематической модели NMF KL), и модель, не использующую их. Результаты сравнения для обоих рассматриваемых корпусов приведены в табл. 8 (комбинирование признаков осуществлялось с помощью логистической регрессии из библиотеки Weka [30]).

Корпус	**Средняя точность**	
	Без тематических признаков	С тематическими признаками
Русский	54.6	**56.3**
Английский	50.4	**51.4**

Таблица 8: Результаты сравнения моделей с тематическими признаками и без них

Мы считаем, что данные результаты, показанные на двух разных коллекциях, подтверждают, что тематические модели действительно вносят дополнительную информацию в процесс автоматического извлечения терминов.

В заключение в таблице 9 представлены первые 10 элементов из списков извлечённых слов-кандидатов, полученных с помощью моделей, учитывающих тематические признаки (при этом термины выделены курсивом).

7 Заключение

В статье представлены результаты экспериментального исследования возможности применения тематических моделей для улучшения качества автоматического извлечения однословных терминов.

Были исследованы различные тематические модели (как вероятностные, так и традиционные методы кластеризации) и предложены несколько модификаций известных признаков для упорядочивания слов-кандидатов по убыванию их терминологичности. В качестве текстовых коллекций бы-

№	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

Таблица 9: Примеры извлечённых слов-кандидатов

ли взяты два различных корпуса: электронные банковские статьи на русском языке и речи с заседаний Европарламента на английском языке.

Эксперименты показали, что независимо от предметной области и языка использование тематической информации способно значительно улучшить качество автоматического извлечения однословных терминов.

Список литературы

[1] K. Ahmad, L. Gillam, L. Tostevin. University of Survey Participation in Trec8. Weirdness indexing for logical document extrapolation and retrieval. In the Proceedings of TREC 1999, 1999.

[2] R. Basili, A. Moschitti, M. Pazienza, F. Zanzotto. A Contrastive Approach to Term Extraction. In the Proceedings of the 4th Terminology and Artificial Intelligence Conference, 2001.

[3] D. Blei and J. Lafferty. Topic Models. Text Mining: Classification, Clustering and Applications, Chapman & Hall, pp. 71–89, 2009.

[4] D. Blei, A. Ng and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, No 3, pp. 993–1022, 2003.

[5] British National Corpus. `http://www.natcorp.ox.ac.uk/`

[6] K. Church and W. Gale. Inverse Document Frequency IDF. A Measure of Deviation from Poisson. In the Proceedings of the Third Workshop on Very Large Corpora. MIT Press, pp. 121–130, 1995.

[7] Chris Ding, Tao Li, Wei Peng. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. Computational Statistics and Data Analysis, No 52, pp. 3913–3927, 2008.

[8] European Parliament Proceedings Parallel Corpus 1996–2011. `http://www.statmt.org/europarl/`

[9] EuroVoc. Multilingual Thesaurus of the European Union. `http://eurovoc.europa.eu/drupal/`

[10] K. Frantzi and S. Ananiadou. Automatic Term Recognition Using Contextual Cues. In the Proceedings of the IJCAI Workshop on Computational Terminology, pp. 29–35, 2002.

[11] A. Gelbukh, G. Sidorov, E. Lavin-Villa, L. Chanona-Hernandez. Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpora. In the Proceedings of the Natural Language Processing and Information Systems, pp. 248–255, 2010.

[12] Q. He, K. Chang, E. Lim, A. Banerjee. Keep It Smile with Time: A Reeximanation of Probabilistic Topic Detection Models. In the Proceedings of IEEE Transactions Pattern Analysis and Machine Intelligence. Volume 32, Issue 10, pp. 1795–1808, 2010.

[13] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In the Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, ACM New York, USA, pp. 50–57, 1999.

[14] S. C. Johnson. Hierarchical Clustering Schemes. Psychometrica, No 2, pp. 241–254, 1967.

[15] D. Kurz and F. Xu. Text Mining for the Extraction of Domain Retrieval Terms and Term Collocations. In the Proceedings of the International Workshop on Computational Approaches to Collocations, 2002.

[16] Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In the Proceedings of NIPS, pp. 556–562, 2000.

[17] N. Loukachevitch. Automatic Term Recognition Needs Multiple Evidence. In the Proceedings of the 8th International Conference on LREC, 2012.

[18] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. In the Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297, 1967.

[19] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[20] H. Nakagawa and T. Mori. A Simple but Powerful Automatic Term Extraction Mehod.

In the Proceedings of the Second International Workshop on Computational Terminology, pp. 29–35, 2002.

[21] H. Nakagawa and T. Mori. Automatic Term Recognition based on Statistics of Compound Nouns and their Components. Terminology, vol. 9, no. 2, pp. 201–219, 2003.

[22] R. Navigli and P. Velardi. Semantic Interpretation of Terminological Strings. In the Proceedings of the 6th International Conference on Terminology and Knowledge Engineering, Springer, pp. 95–100, 2002.

[23] M. A. Nokel, E. I. Bolshakova, N. V. Loukachevitch. Combining Multiple Features for Single-Word Term Extraction. Компьютерная лингвистика и интеллектуальные технологии. По материалам конференции Диалог-2012, Бекасово, pp. 490–501.

[24] Y. Park, R. J. Bird, B. Boguraev. Automatic glossary extraction beyond terminology identification. In the Proceedings of the 19th International Conference on Computational Linguistics, 2002.

[25] P. Pecina and P. Schlesinger. Combining Association Measures for Collocation Extraction. In the Proceedings of the COLING/ACL, ACL Press, pp. 651–658, 2006.

[26] A. Peñas, V. Verdejo, J. Gonzalo. Corbus-based Terminology Extraction Applied to Information Access. In the Proceedings of the Corpus Linguistics 2001 Conference, pp. 458–465, 2001.

[27] X.-H. Phan, C.-T. Nguyen. GibbsLDA++: A C/C++ implementation of latent Dirichlet Allocation (LDA), 2007.

[28] G. Salton. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, 1989.

[29] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler. Exploring Topic Coherence over many models and many topics. In the Proceedings of EMNLP-CoNLL, pp. 952–961, 2012.

[30] Weka 3. Data Mining Software in Java. `http://www.cs.waikato.ac.nz/ml/weka`

[31] W. Wong, W. Liu, M. Bennamoun. Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency. In the Proceedings of the 6th Australasian Conference on Data Mining, pp. 47–54, 2007.

[32] W. Xu, X. Liu, Y. Gong. Document Clustering Based On Non-negative Matrix Factorization. In the Proceedings of SIRGIR, pp. 267–273, 2003.

[33] Shi Zhong. Efficient Online Spherical K-means Clustering. In the Proceedings of IEEE-IJCNN, Monreal, Canada, July 31 – August 4, pp. 3180–3185, 2005.

[34] К. В. Воронцов и А. А. Потапенко. Регуляризация, робастность и разреженность вероятностных тематических моделей. Журнал “Компьютерные исследования и моделирование”, т. 4, №12, с. 693–706, 2012.

[35] Н. В. Лукашевич. Тезаурусы в задачах информационного поиска. Москва: Издательство Московского университета, 2011.

Application of Topic Models to the Task of Single-Word Term Extraction

Michael Nokel, Natalia Loukachevitch

The paper describes the results of an experimental study of statistical topic models applied to the task of single-word term extraction. The English part of the Europarl corpus and the Russian articles taken from online banking magazines were used as target text collections. The experiments demonstrate that topic information significantly improves the quality of single-word term extraction, regardless of the subject area and the language used.