УДК 004.852

ТЕМАТИЧЕСКИЕ МОДЕЛИ: ДОБАВЛЕНИЕ БИГРАММ И УЧЕТ СХОДСТВА МЕЖДУ УНИГРАММАМИ И БИГРАММАМИ

М. А. Нокель[1], **Н. В. Лукашевич**[2]

Представлены результаты экспериментов по добавлению биграмм в тематические модели и учету сходства между ними и униграммами. Предложен новый алгоритм PLSA-SIM, являющийся модификацией алгоритма построения тематических моделей PLSA (Probabilistic Latent Semantic Analysis). Предложенный алгоритм позволяет добавлять биграммы и учитывать сходство между ними и униграммными компонентами. Исследована возможность применения ассоциативных мер для выбора и последующего включения биграмм в тематические модели. В качестве текстовых коллекций взяты русскоязычная подборка статей из электронных банковских журналов, английские части корпусов параллельных текстов Europarl и JRC-Acquiz и англоязычный архив исследовательских работ по компьютерной лингвистике ACL Anthology. Выполненные эксперименты показывают, что существует подгруппа тестируемых мер, упорядочивающих биграммы таким образом, что при последующем их добавлении в предложенный алгоритм PLSA-SIM качество получающихся тематических моделей значительно повышается. Предложен новый итеративный алгоритм PLSA-ITER без учителя, позволяющий добавлять наиболее подходящие биграммы. Эксперименты показывают дальнейшее улучшение качества тематических моделей по сравнению с исходным алгоритмом PLSA.

Ключевые слова: тематические модели, PLSA (Probabilistic Latent Semantic Analysis), ассоциативные меры, биграммы, согласованность тем, перплексия.

1. Введение. *Вероятностное тематическое моделирование* (далее просто *тематическое моделирование*) — одна из активно развивающихся областей статистического анализа текстов. Тематические модели предназначены для выявления скрытых тем в текстовых коллекциях. Они определяют, какие темы присутствуют в каждом документе коллекции и какие слова задают каждую такую тему. При этом темы представляются в виде дискретных распределений на множестве слов, а документы — в виде дискретных распределений на множестве тем [1]. Например, в коллекциях исследовательских работ темы будут соответствовать теориям, методам и алгоритмам, при описании которых используется устоявшаяся терминология. В коллекциях новостей темы могут соответствовать событиям, компаниям, различным деятелям и т.д. Примером такой темы может служить следующая (для наглядности вероятности слов опущены): *денежный, деньги, обращение, масса, факторинг, средство, функция, оборот, факторинговый, товар и др.*

Тематические модели осуществляют нечеткую кластеризацию слов и документов по кластерам-темам. Это означает, что слово или документ могут относиться к нескольким темам с различными вероятностями. При этом слова, встречающиеся в одних и тех же контекстах, с большей вероятностью попадут в одну и ту же тему, а слова, употребляющиеся в различных контекстах, распределятся между разными темами.

На данный момент тематические модели успешно применяются в различных задачах информационного поиска [2], разрешения морфологической неоднозначности [3], многодокументного аннотирования [4], машинного перевода [5], категоризации и кластеризации документов [6]. Кроме того, на их основе были достигнуты значительные успехи в выявлении трендов в научных публикациях и новостных потоках [7], обработке аудио- и видеосигналов [8] и многих других задачах.

Самыми широко распространенными алгоритмами построения тематических моделей являются метод вероятностного латентного семантического анализа PLSA (Probabilistic Latent Semantic Analysis) [9], не связанный ни с какими параметрическими априорными распределениями, и метод латентного размещения Дирихле (LDA, Latent Dirichlet Allocation) [1], использующий априорное распределение Дирихле. Кроме того, существует множество их обобщений и модификаций [10]. Однако многие проблемы, в частности

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 119991, Москва; аспирант, e-mail: mnokel@gmail.com

[2] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, Ленинские горы, 119992, Москва; вед. науч. сотр., e-mail: louk_nat@mail.ru

проблемы слабой интерпретируемости выделяемых тем, на данный момент не имеют окончательного решения.

Предполагается, что пользователь, увидев список самых характерных для данной темы слов, сможет понять, о чем эта тема, дать ей уникальное название. Однако темы, найденные с помощью той или иной тематической модели, зачастую оказываются непонятными, содержащими много слабосвязанных между собой слов. Одним из основных недостатков тематических моделей, приводящих к описанной выше проблеме, является использование модели "мешка слов", в которой каждый документ представляется в виде множества встречающихся в нем слов. Данная модель никак не учитывает порядок слов и основывается на гипотезе независимости слов друг от друга в тексте. Это предположение оправдано с точки зрения вычислительной эффективности, но оно далеко от реальности. Так, некоторые слова меняют свой смысл при объединении в словосочетания: например, биграмма *"точка зрения"* плохо связана со своими униграммными компонентами *"точка"* и *"зрение"*.

Следует отметить, что на данный момент проведено достаточно большое количество исследований, посвященных изучению проблемы добавления словосочетаний, n-грамм и многословных выражений в тематические модели. Однако зачастую это приводит к ухудшению качества модели в связи с увеличением размера словаря или к значительному усложнению модели [11–13].

В настоящей статье предлагается новый метод, позволяющий добавить биграммы в тематические модели и сохранить связь между ними и униграммными компонентами (такими как *бухгалтерский — бухгалтерская отчетность — бухгалтерский учет* и *бюджетный — бюджетный кодекс — бюджетная политика — бюджетные расходы — бюджетная система*). Предложенный подход рассматривает биграммы уже не как "черные ящики", а учитывает взаимосвязь между ними и униграммными компонентами.

Все эксперименты, описанные в этой статье, проведены на основе алгоритма PLSA и его модификаций на четырех текстовых коллекциях разных языков и предметных областей: английской части корпуса параллельных текстов Europarl (`http://www.statmt.org/europarl`), английской части корпуса параллельных текстов JRC-Acquiz (`http://ipsc.jrc.ec.europa.eu/index.php?id=198`), архиве исследовательских работ по компьютерной лингвистике ACL Anthology (`http://acl-arc.comp.nus.edu.sg/`) и подборке статей из русскоязычных банковских электронных журналов.

Предложенные алгоритмы показывают улучшение характеристик качества тематических моделей, включая интерпретацию тем экспертами. Отдельно отметим, что предлагаемая модификация была также применена и к алгоритму LDA и результаты оказались аналогичны. Однако в целях экономии места в статье представлены результаты экспериментов только с применением алгоритма PLSA.

Статья организована следующим образом. В разделе 2 рассматриваются близкие работы. В разделе 3 предлагается новый алгоритм PLSA-SIM, являющийся модификацией исходного алгоритма PLSA с добавлением биграмм и учета сходства между ними и униграммными компонентами. В разделе 4 предлагается итеративный алгоритм без учителя PLSA-ITER, позволяющий добавлять наиболее подходящие биграммы. В разделе 5 описываются все текстовые коллекции, использующиеся в экспериментах, все стадии их предобработки и метрики оценки качества тематических моделей. В разделе 6 приводится обширный анализ различных мер упорядочивания биграмм для последующего их включения в тематические модели. И наконец, в последнем разделе формулируются выводы.

2. Близкие работы.

2.1. Тематические модели. На сегодняшний день разработано достаточно большое количество алгоритмов построения тематических моделей. Исторически первыми появились методы, основанные на традиционной кластеризации текстов [14]. Они основываются на методах "жесткой" кластеризации, рассматривающих каждый документ как разреженный вектор в пространстве слов большой размерности [15]. После окончания работы алгоритма кластеризации каждый получившийся кластер рассматривается как отдельная тема, содержащая в себе слова с вероятностями, вычисленными по следующей формуле:

$$P\big(w|t\big) = \frac{f\big(w|t\big)}{\sum\limits_{w \in t} f\big(w|t\big)},$$

где $f\big(w|t\big)$ — частотность слова w в кластере-теме t.

В качестве алгоритмов кластеризации могут выступать любые известные методы: K-средних [16] и его модификации, различные алгоритмы иерархической кластеризации и др. Естественным ограничением подобных моделей является отнесение документа лишь к одной теме, в то время как очевидно, что практически в любом документе затрагивается несколько различных тем.

В последнее время появились и стали активно использоваться вероятностные методы выявления тем в документах, рассматривающие каждый отдельный текст в виде смеси нескольких тем, а каждую тему — в виде некоторого вероятностного распределения над словами. При этом порождение слов происходит по правилу

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

где $P\big(w|t\big)$ и $P\big(t|d\big)$ — скрытые распределения слов по темам и тем по документам, а $P\big(w|d\big)$ — наблюдаемое распределение слов по документам.

При известных распределениях $P\big(w|t\big)$ и $P\big(t|d\big)$ порождение слов в документах происходит согласно следующему алгоритму.

Алгоритм 1. Порождение коллекции текстов с помощью тематической модели.

Вход: распределения $P\big(w|t\big)$ и $P\big(t|d\big)$
Выход: коллекция документов D
for $d \in D$ **do**
    Задать длину документа n_d
    for $i = 1, \ldots, n_d$ **do**
        Сэмплировать тему t из распределения $P\big(t|d\big)$
        Сэмплировать слово w из распределения $P\big(w|t\big)$
        Добавить в документ d коллекции D слово w

Задача построения тематической модели состоит в восстановлении скрытых распределений $P\big(w|t\big)$ и $P\big(t|d\big)$ по известной коллекции D. Для ее решения используется метод максимума правдоподобия:

$$\ln \prod_{d\in D}\prod_{w\in d} P\big(w|d\big)^{n_{dw}} = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} \longrightarrow \max_{\Phi,\Theta}$$

при ограничениях неотрицательности и нормировки:

$$\phi_{wt} = P\big(w|t\big) \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \quad \theta_{td} = P\big(t|d\big) \geqslant 0; \quad \sum_{d\in D} \theta_{td} = 1.$$

Здесь n_{dw} — частотность слова w в документе d, D — коллекция документов, T — множество выделяемых тем в коллекции D, W — множество уникальных слов в коллекции D (словарь коллекции D), $\Phi = \big\{\phi_{wt}\big\}_{W\times T} = \Big\{P\big(w|t\big)\Big\}_{W\times T}$ и $\Theta = \big\{\theta_{td}\big\}_{T\times D} = \Big\{P\big(t|d\big)\Big\}_{T\times D}$ — матрицы скрытых распределений $P\big(w|t\big)$ и $P\big(t|d\big)$ соответственно.

Тематическая модель зависит от нескольких скрытых переменных; для нахождения оценок максимального правдоподобия параметров Φ и Θ используется EM-алгоритм (Expectation–Maximization) [17]. Это итеративный алгоритм, каждая итерация которого состоит из двух шагов:

— E-шаг (Expectation-шаг). Вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. В рассматриваемой задаче условные вероятности $P\big(t|d,w\big)$ для всех тем t, документов d и слов w вычисляются через скрытые параметры ϕ_{wt} и θ_{td} по формуле Байеса

$$P\big(t|d,w\big) = \frac{P\big(w,t|d\big)}{P\big(w|d\big)} = \frac{P\big(w|t\big)P\big(t|d\big)}{P\big(w|d\big)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}};$$

— M-шаг (Maximization-шаг). Находится оценка максимального правдоподобия, тем самым увеличивается ожидаемое правдоподобие. В рассматриваемой задаче частотные оценки условных вероятностей вычисляются путем суммирования счетчика $n_{dwt} = n_{dw}P\big(t|d,w\big)$, показывающего, сколько раз слово w в документе d отнеслось к теме t:

$$\phi_{wt} = \frac{n_{wt}}{n_t} = \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}; \quad \theta_{td} = \frac{n_{td}}{n_d} = \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}}.$$

Данные итерации повторяются до сходимости.

Самыми известными представителями этой категории моделей являются метод LDA [1], использующий априорное распределение параметров Дирихле, и метод PLSA [9], не использующий никаких априорных распределений. Известно, что эти алгоритмы можно рассматривать как EM-подобные алгоритмы [18], различающиеся порядком применения формул E-шага и M-шага, модификациями M-шага и способами распределения частотности n_{dw} по темам.

2.2. Добавление словосочетаний в тематические модели. Все описанные в разделе 2.1 тематические модели работают только со словами, основываясь на гипотезе "мешка слов" об их условной независимости друг от друга. Следует отметить работы, в которых исследуется вопрос добавления словосочетаний в тематические модели для улучшения их качества [11–13]. На данный момент существуют два основных подхода к решению этой проблемы: создание единой унифицированной вероятностной тематической модели, построенной с учетом слов и словосочетаний, и предварительное извлечение словосочетаний для последующего добавления в тематические модели.

Большинство существующих работ посвящено именно первому подходу. Так, первая попытка уйти от гипотезы "мешка слов" была предпринята в работе [11], в которой предложена биграммная тематическая модель (БТМ). В этой модели вводится понятие "порядка слов", и вероятность появления слова в тексте зависит от непосредственно предшествующего слова. При этом БТМ работает только с биграммами, и никакие слова не рассматриваются как униграммы. Модель словосочетаний LDA, представленная в работе [12], расширяет биграммную тематическую модель за счет введения дополнительных переменных, которые для каждого слова в документе указывают, составляет ли данное слово с предыдущим словом биграмму или нет. При этом темы по-прежнему задаются отдельными униграммами, а биграммы задаются вероятностями перехода от слова к слову вне зависимости от темы. В работе [13] представлена N-граммная тематическая модель, усложняющая обе предыдущие модели для обеспечения возможности формирования биграмм в текстах в зависимости от контекста. В работе [19] предложена тематическая модель слово–символ, уходящая от использовавшегося во всех предыдущих моделях предположения о том, что тема каждой N-граммы определяется в зависимости от тем слов, образующих словосочетание. Эта модель изначально разрабатывалась и оказалась наиболее пригодной для китайского языка. В работе [20] устанавливается связь между LDA и вероятностными контекстно-свободными грамматиками и предлагаются две новые вероятностные модели, сочетающие в себе идеи из LDA и вероятностных контекстно-свободных грамматик для добавления словосочетаний и имен собственных в тематические модели.

Несмотря на то что все описанные выше тематические модели имеют теоретически элегантное обоснование, у них очень много параметров для настройки, что делает их интересными только с теоретической точки зрения и ограничивает возможность их применения на практике на реальных данных. Так, число параметров у биграммной тематической модели равно W^2T, у модели словосочетаний LDA — $W^2T + W^2$, у N-граммной тематической модели — W^NT, в то время как у LDA — WT, у PLSA — $WT + DT$, где W — размер словаря (т.е. количество уникальных слов и словосочетаний в коллекции), D — количество документов в текстовой коллекции, T — количество тем и N — максимальный размер N-грамм в модели.

Ко второму типу методов, добавляющих словосочетания в тематические модели, относится алгоритм, предложенный в работе [21]. На этапе предобработки авторы извлекают все встретившиеся в коллекции текстов биграммы, после чего упорядочивают их в соответствии с мерой t-теста. Затем в документах заменяют отдельные униграммы лучшими по данной мере биграммами, добавляя их в словарь коллекции (авторы рассматривают 1000 лучших биграмм). При этом для оценки качества тематических моделей используются 2 метрики: перплексия, являющаяся стандартной метрикой качества всех языковых моделей, и согласованность тем [22]. Показано, что добавление биграмм в тематические модели приводит к ухудшению перплексии, но к улучшению согласованности тем.

Настоящая работа тоже относится ко второму типу методов и отличается от работы [21] тем, что описываемый здесь подход рассматривает биграммы не как "черные ящики" без связей с остальными словами, а учитывает внутреннюю структуру биграмм и взаимосвязь между ними и униграммными компонентами, что приводит к улучшению обоих показателей: и перплексии, и согласованности тем.

Идея использования априорных лингвистических знаний в тематических моделях сама по себе не нова. Так, в работе [23] предметно-ориентированные знания представляются в виде Must-link и Cannot-link примитивов с помощью априорного леса Дирихле. Эти примитивы отвечают за то, чтобы слова порождались одними и теми же или разными темами. Однако впоследствии было замечено, что данный метод может привести к экспоненциальному росту при кодировании Cannot-link примитивов, а потому его сложно применять на реальных данных с большим количеством ограничений [24]. Другой способ добавления подобных знаний представлен в работе [25], в которой был предложен частично обучаемый

с учителем EM-алгоритм для группировки выражений в предопределенные пользователем категории. Для обеспечения наилучшей инициализации EM-алгоритма авторы предлагают использовать априорное знание о том, что синонимы и выражения, содержащие в себе одинаковые слова, должны, скорее всего, относиться к одним и тем же группам. Наша работа отличается от рассмотренных выше тем, что в ней сходства между биграммами и униграммами добавляются в тематическую модель естественным образом путем подсчета их совместной встречаемости в документах текстовой коллекции. Предлагаемый подход тоже не изменяет число параметров модели по сравнению с оригинальным алгоритмом PLSA.

3. Алгоритм PLSA-SIM. Тематические модели PLSA и LDA используют модель "мешка слов", не учитывающую порядок слов и предполагающую независимость появлений слов в документах. Более того, биграммы обычно тоже добавляются в тематические модели как "черные ящики" без всяких связей с остальными словами. Процесс добавления биграмм получается следующий. Вначале биграммы добавляются в словарь коллекции, после чего в каждом документе униграммные компоненты добавляемых биграмм заменяются этими биграммами [21]. Таким образом, предположение модели "мешка слов" выполняется.

Данное предположение упрощает выкладки, но далеко от реальности, поскольку в документах есть много слов и словосочетаний, связанных между собой по смыслу: например, биграммы и униграммы, содержащие одно общее слово: *бюджетный — бюджетные расходы — бюджетные доходы — бюджетные средства* или *жилищный — жилищная ипотека — жилищный кредит — жилищное кредитование — жилищный рынок — жилищное строительство* и др. Следует отметить, что у таких биграмм не только есть одинаковые слова, но многие из них обладают также семантической и тематической близостью. В то же время у других биграмм, содержащих общие слова (например, у идиом), могут быть значительные семантические различия. Для того чтобы учесть эти различные ситуации, было выдвинуто предположение, что похожие биграммы, содержащие общие униграммные компоненты, должны часто относиться к одним и тем же темам, если они часто встречаются вместе в рамках одних и тех же документов.

Для проверки данной гипотезы были составлены множества похожих униграмм и биграмм, содержащих общие слова, и предложен новый алгоритм PLSA-SIM, являющийся модификацией исходного алгоритма PLSA. При описании проведенной модификации будет использоваться описание алгоритма PLSA, представленное в работе [26], и следующие обозначения:

D — коллекция документов;

T — множество полученных тем;

W — словарь коллекции (множество уникальных униграмм и биграмм в коллекции документов D);

$\Phi = \left\{\phi_{wt} = P\big(w|t\big)\right\}$ — распределение униграмм (биграмм) w по темам t;

$\Theta = \left\{\theta_{td} = P\big(t|d\big)\right\}$ — распределение тем t по документам d;

$S = \left\{S_w\right\}$ — множество похожих униграмм и биграмм, где S_w — множество униграмм и биграмм, похожих на w:

$$S_w = \left\{w, \bigcup_v wv, \bigcup_v vw\right\},$$

где w — лемматизированная униграмма, wv и vw — лемматизированные биграммы, содержащие v. В табл. 1 приведены несколько примеров множеств похожих униграмм и биграмм вместе с центральной униграммой в каждом множестве;

n_{dw} и n_{ds} — частотности униграмм (биграмм) w и s в документе d;

n_d — длина документа d (общее число униграмм (биграмм) в документе d);

$\widehat{n}_{wt}$ — оценка частотности униграммы (биграммы) w в теме t;

$\widehat{n}_{td}$ — оценка частотности темы t в документе d;

$\widehat{n}_t$ — оценка частотности темы t в коллекции документов D;

$P\big(t|d,w\big)$ — условная вероятность отнесения вхождения униграммы (биграммы) w в документ d к теме t.

При описании алгоритма PLSA-SIM (см. алгоритм 2) за основу был взят EM-алгоритм для модели PLSA с внесенным E-шагом внутрь M-шага для избежания хранения трехмерного массива $P(t|d,w)$ (см. раздел 2.1). Единственная модификация касается строчки 7, где в рассмотрение добавляются предварительно вычисленные множества похожих слов и биграмм. Тем самым вес подобных слов увеличивается в каждом документе текстовой коллекции.

Более формально: если похожие униграммы и биграммы встречаются вместе в рамках одного и того же документа d, то предложенный алгоритм старается их отнести к одним и тем же темам, предполагая, что такие униграммы и биграммы обладают семантической и тематической близостью. Однако если же униграммы и биграммы из одного и того же множества S_w не встречаются вместе в одном и том же документе, то исходный алгоритм PLSA не модифицируется. Предполагается, что такие униграммы и биграммы обладают семантическими различиями.

Таблица 1

Примеры множеств похожих униграмм и биграмм в алгоритме PLSA-SIM

Множество похожих слов и биграмм	Центральная униграмма
Кластеризация; *кластеризация* документов; *кластеризация* текстов	*Кластеризация*
Недвижимость; рынок *недвижимости*; строительство *недвижимости*	*Недвижимость*
Мобильный; *мобильное* устройство; *мобильный* телефон	*Мобильный*

Отдельно следует отметить, что предлагаемая модификация никак не увеличивает число параметров оригинального алгоритма PLSA — оно остается таким же — равным $WT + DT$ (см. раздел 2.2).

Алгоритм 2. Алгоритм PLSA-SIM: PLSA с похожими униграммами и биграммами.

Вход: коллекция документов D, количество тем T, начальные приближения Φ и Θ, множества похожих униграмм и биграмм S

Выход: распределения Φ и Θ

while не выполнится критерий остановки **do**

for $d \in D$, $w \in W$, $t \in T$ **do**

$\widehat{n}_{wt} = 0$, $\widehat{n}_{td} = 0$, $\widehat{n}_t = 0$

for $d \in D$, $w \in W$ **do**

for $t \in T$ **do**

$$P(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}$$

$$\widehat{n}_{wt}, \widehat{n}_{td}, \widehat{n}_t += \left(n_{dw} + \sum_{s \in S_w} n_{ds}\right) P(t|d,w)$$

for $d \in D$, $w \in W$ **do**

$$\phi_{wt} = \frac{\widehat{n}_{wt}}{\widehat{n}_t}$$

for $d \in D$, $t \in T$ **do**

$$\theta_{td} = \frac{\widehat{n}_{td}}{n_d}$$

4. Итеративный алгоритм PLSA-ITER. Было сделано предположение, что для добавления в тематические модели можно выбирать наиболее подходящие биграммы исходя из вида верхушек списков униграмм, образующих эти темы. Для этой цели можно попытаться составить в каждой теме из первых униграмм все возможные биграммы, которые впоследствии следует добавить в тематическую модель при обучении. Так, например, если в какой-то теме в верхней части списка окажутся униграммы *"ценный"* и *"бумага"*, то можно попытаться добавить в тематическую модель лемматизированную биграмму *"ценный бумага"*. Для проверки данной гипотезы был предложен новый итеративный алгоритм выбора биграмм PLSA-ITER.

При описании предлагаемого алгоритма используются следующие дополнительные обозначения:

B — множество всех биграмм в коллекции документов D;

B_A — множество биграмм, добавленных в тематическую модель;

B_i — множество биграмм, добавленных в тематическую модель на i-й итерации;

S_i — множество потенциальных кандидатов на похожие униграммы и биграммы на i-й итерации;

$\left(u_1^t, u_2^t, \ldots, u_{10}^t\right)$ — первые 10 униграмм в теме t;

$TF\left(u_i^t, u_j^t\right)$ — частотность биграммы $\left(u_i^t, u_j^t\right)$.

На каждой итерации алгоритм PLSA-ITER (см. алгоритм 3) добавляет в множество кандидатов в похожие униграммы и биграммы первые 10 униграмм из каждой темы. Кроме того, в это множество и в саму тематическую модель добавляются все биграммы, которые могут быть образованы с помощью этих первых 10 униграмм. При этом если из пары униграмм можно составить две различные биграммы, то рассматривается только наиболее частотная биграмма. Было принято решение анализировать только первые 10 униграмм в темах, поскольку одной из целевых метрик качества является согласованность тем, использующая именно это множество (см. раздел 5.2). В соответствии с данным алгоритмом в тематические модели добавляются только те биграммы, которые образуются с помощью первых 10 униграмм в темах, а такие биграммы с большой вероятностью могут оказаться наиболее подходящими.

Алгоритм 3. Итеративный алгоритм PLSA-ITER.

Вход: коллекция документов D, число тем $|T|$, множество биграмм B
Выход: полученные темы T
Запуск оригинального алгоритма PLSA на коллекции документов D для получения тем T
$B_A = \emptyset$
while не выполнится критерий остановки **do**
    $S_i = \emptyset$, $B_i = \emptyset$
    for $t \in T$ **do**
        $S_i = S_i \bigcup \left\{u_1^t, u_2^t, \ldots, u_{10}^t\right\}$
        for $u_i^t, u_j^t \in \left(u_1^t, u_2^t, \ldots, u_{10}^t\right)$ **do**
            if $\left(u_i^t, u_j^t\right) \in B$ and $\left(u_j^t, u_i^t\right) \in B$ and $TF\left(u_i^t, u_j^t\right) > TF\left(u_j^t, u_i^t\right)$ **then**
                $B_i = B_i \bigcup \left\{\left(u_i^t, u_j^t\right)\right\}$
    Формирование множества похожих униграмм и биграмм S из $S_i \bigcup B_i$
    $B_A = B_A \bigcup B_i$
    Запуск PLSA-SIM с множеством похожих униграмм и биграмм S и множеством биграмм B

Следует отметить, что число параметров итеративного алгоритма PLSA-ITER остается таким же, как и в случае с оригинальным алгоритмом PLSA — равным $WT + DT$.

5. Текстовые коллекции и методы оценки качества.

5.1. Текстовые коллекции и предобработка. В экспериментах, описанных в настоящей статье, использовались текстовые коллекции различных языков и предметных областей:

— для английской части исследования были выбраны три различные текстовые коллекции:

  а) многоязычный корпус параллельных текстов Europarl, составленный из протоколов заседаний Европарламента (`http://www.statmt.org/europarl`); английская часть данного корпуса содержит примерно 54 миллиона слов в 9672 документах;

  б) многоязычный корпус параллельных текстов JRC-Acquiz, составленный из ряда статей законодательства Евросоюза с 1950 по 2005 г.г. (`http://ipsc.jrc.ec.europa.eu/index.php?id=198`); английская часть данного корпуса содержит примерно 45 миллионов слов в 23 545 документах;

  в) архив работ по компьютерной лингвистике ACL Anthology (`http://acl-arc.comp.nus.edu.sg/`); в данном корпусе содержится примерно 42 миллиона слов в 10 921 документе;

— для русской части исследования была взята подборка статей из электронных банковских журналов (таких как “Аудитор”, РБК, “Банки и Технологии” и др.); в данной коллекции оказалось примерно 18.5 миллионов слов в 10 422 документах.

При предобработке текстов проведен их морфологический анализ. Для английских корпусов текстов использованы средства Stanford CoreNLP (`http://nlp.stanford.edu/software/corenlp.shtml`), а для русского корпуса — собственный морфологический анализатор. Все слова были лемматизированы, т.е.

приведены к начальной форме. В качестве слов, участвующих в образовании тем, рассматривались только *прилагательные*, *существительные*, *наречия* и *глаголы*, поскольку остальные служебные слова не играют особой роли в данном процессе. Слова, встретившиеся в каждой из текстовых коллекций меньше 5 раз, исключались из рассмотрения. Кроме того, были извлечены все встретившиеся в коллекциях биграммы в формах:

— *существительное* + *существительное*, *прилагательное* + *существительное*, *существительное* + *of* + *существительное* — для английских текстовых коллекций;

— *существительное* + *существительное в родительном падеже*, *прилагательное* + *существительное* — для русской текстовой коллекции.

В нашей работе рассматривались только такие биграммы, поскольку темы, как правило, образуются с помощью существительных и именных групп [13].

5.2. Методы оценки качества тематических моделей. Оценка качества тематических моделей является сложной проблемой. В отличие от задач классификации здесь нет четкого понятия "ошибки". В рамках нашей работы мы использовали 4 различные метрики оценки качества тематических моделей.

Наиболее широко известным критерием является *перплексия*, используемая для оценки языковых моделей в компьютерной лингвистике [10]. Это мера несоответствия модели $p\big(w|d\big)$ словам w, наблюдаемым в документах коллекции, определяется через логарифм правдоподобия:

$$\text{Perplexity}\,(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p\big(w|d\big)\right),$$

где n — число всех рассматриваемых слов в коллекции, D — множество всех документов в коллекции, n_{dw} — частотность слова w в документе d и $p\big(w|d\big)$ — вероятность появления слова w в документе d.

Чем меньше значение перплексии, тем лучше модель предсказывает появление слов w в документах текстовой коллекции D. Поскольку известно, что перплексия, вычисленная по той же самой обучающей коллекции документов, склонна к переобучению и может давать оптимистически заниженные оценки [1], в предлагаемой работе используется стандартный метод вычисления контрольной перплексии, описанный в работе [18]. Текстовая коллекция изначально разбивается на 2 части: обучающую D, по которой строится модель, и контрольную D', по которой вычисляется данная метрика. В экспериментах, описанных в этой статье, использовалось случайное разбиение коллекции в перплексии $|D| : |D'| = 9 : 1$. Параметры ϕ_{wt} оценивались только по обучающей коллекции D. После обучения параметры ϕ_{wt} фиксировались, а каждый контрольный документ $d \in D'$ разбивался случайным образом на 2 половины: по первой половине оценивались параметры θ_{td}, а по второй вычислялась контрольная перплексия. При этом новые слова, которые ни разу не встретились во всей обучающей коллекции D, но попавшие во вторую половину контрольных документов, игнорировались. Хотя на данный момент существуют работы, утверждающие, что перплексию нецелесообразно использовать для оценки качества тематических моделей [27], данная метрика по-прежнему широко используется для сравнения различных тематических моделей между собой.

Другим традиционным способом оценки качества тематических моделей, тоже использующимся в рамках нашей работы, являются *экспертные оценки*. Экспертам были предоставлены полученные темы в виде списков слов и словосочетаний, упорядоченных по убыванию степени принадлежности, и им было предложено решить, является ли каждая из предоставленных тем в какой-то степени последовательной, осмысленной и интерпретируемой. Индикатором такой темы служит возможность дать ей некоторое обобщенное название. Таким образом, перед экспертами-лингвистами была поставлена задача классифицировать все предоставленные им темы на две категории в зависимости от того, можно ли дать теме некоторое название или нет. В последнем случае это означало бы, что тема состоит из разрозненных слов, не связанных между собой по смыслу, и является "мусорной". В табл. 2 приведены примеры согласованной темы с названием, данным экспертами, и несогласованной темы, которой невозможно дать никакого обобщенного названия.

Поскольку экспертная оценка полученных тем является дорогостоящей операцией и требует значительного времени, за последнее время было предпринято несколько попыток предложить способ автоматической оценки качества тематических моделей, который был бы никак не связан с перплексией и коррелировал бы с мнениями экспертов. Данная постановка задачи является очень сложной, поскольку эксперты могут достаточно сильно расходиться во мнениях. Однако в недавних работах [22, 28] было показано, что можно автоматически оценивать согласованность тем, основываясь на семантике слов, с точностью, почти совпадающей с экспертными оценками. Предложенная метрика измеряет интерпретируемость тем, основываясь на способах ручной оценки экспертом [22]. Поскольку темы, как правило,

предоставляются экспертам для проверки в виде первых N слов, согласованность тем оценивает то, насколько данные слова соответствуют рассматриваемой теме. При этом тема называется *согласованной*, если наиболее частые в этой теме слова неслучайно часто совместно встречаются рядом в документах коллекции. В работе [22] авторы предложили использовать автоматический способ вычисления данной метрики исходя из меры взаимной информации (**TC-PMI**):

$$\text{TC-PMI}(t) = \sum_{j=2}^{10} \sum_{i=1}^{j-1} \ln \frac{P(w_i, w_j)}{P(w_i)P(w_j)},$$

где $(w_1, w_2, \ldots, w_{10})$ — первые 10 слов в рассматриваемой теме t, $P(w_i)$ и $P(w_j)$ — вероятности униграмм w_i и w_j соответственно, а $P(w_i, w_j)$ — вероятность биграммы (w_i, w_j). Итоговая мера согласованности тем вычисляется усреднением TC-PMI(t) по всем темам t.

Таблица 2

Примеры множеств похожих униграмм и биграмм в алгоритме PLSA-SIM

Верхняя часть списка слов из темы	Название темы
Быть, человек, люди, год, когда, время, женщина, жизнь, деньги	—
Предприятие, лизинг, имущество, объект, лизинговый, аренда, земельный	*Лизинг*

Чем выше значение данной меры, тем лучше согласованы выявленные темы. Данная метрика показывает очень высокую корреляцию с оценками экспертов [22]. Она рассматривает только первые 10 слов в каждой теме, поскольку они, как правило, предоставляют достаточно информации для формирования предмета темы и отличительных черт одной темы от другой. Согласованность тем становится все более широко используемым методом оценки качества тематических моделей наряду с перплексией. Так, в работе [29] показано, что данная метрика очень сильно коррелирует с оценками экспертов, а в работе [30] она просто используется как один из способов оценки качества тематических моделей.

В соответствии с подходом, изложенным в работе [28], в нашей работе вероятности униграмм и биграмм вычисляются путем деления количества документов, в которых встретилась та или иная униграмма или биграмма, на число всех документов в коллекции. Во избежание оптимистически завышенных значений согласованности тем и для вычисления подобных вероятностей использовался сторонний корпус текстов — а именно, xml-дампы русской и английской Википедии, как было предложено в статье [22].

Кроме того, существует и другой вариант вычисления меры согласованности тем на основе логарифма от условной вероятности (**TC-LCP**), предложенный в работе [28]. Он оценивает вероятность менее частотного слова при условии более частотного:

$$\text{TC-LCP}(t) = \sum_{j=2}^{10} \sum_{i=1}^{j-1} \ln \frac{P(w_i, w_j)}{P(w_i)}.$$

Данный вариант меры согласованности тем в нашей работе не рассматривается, поскольку в статье [21] было показано, что этот вариант работает значительно хуже, чем TC-PMI.

Следует отметить, что в предлагаемой модификации алгоритма PLSA, добавляющей знаний о похожести слов и словосочетаний в тематические модели, такие слова с большей вероятностью окажутся среди первых десяти в полученных темах. Тем самым происходит неявная максимизация меры TC-PMI, поскольку похожие слова и словосочетания склонны встречаться в рамках одних и тех же документов. Поэтому было принято решение модифицировать данную метрику для учета только первых 10 непохожих слов в каждой теме (в дальнейшем в настоящей статье данная метрика будет обозначаться как **TC-PMI-nSIM**).

6. Интеграция биграмм в тематические модели.

6.1. Интеграция биграмм в тематические модели с помощью алгоритма PLSA-SIM. На первом этапе экспериментов сравнивались результаты работы предложенного алгоритма PLSA-SIM с оригинальным алгоритмом PLSA. Для этой цели были извлечены все биграммы, встретившиеся в коллекции не менее 5 раз. Для последующего упорядочивания извлеченных биграмм применялись ассоциативные меры — математические критерии, определяющие силу связи между составными частями фраз, основываясь на частотах встречаемости отдельных слов и словосочетаний целиком. В экспериментах использовались следующие 16 ассоциативных мер: *взаимная информация (MI)* [31], *дополненная взаимная*

информация (*дополненная MI*) [32], *нормализованная взаимная информация* (*нормализованная MI*) [33], *настоящая взаимная информация* (*настоящая MI*) [34], *кубическая взаимная информация* (*кубическая MI*) [35], *симметричная условная вероятность* [36], *коэффициент Сёренсена (DC)* [37], *модифицированный коэффициент Сёренсена* (*модифицированный DC*) [38], *Gravity Count* [39], *коэффициент простого соответствия*, *коэффициент Юла* [35], *коэффициент Кульчинского* [40], *коэффициент Жаккара* [41], *T-Score*, *Хи-квадрат* [42] и *отношение логарифмического правдоподобия* [43]. В качестве простой меры упорядочивания биграмм ("baseline") была выбрана обычная *частотность* (*TF*).

В соответствии с результатами, представленными в работе [21], в тематические модели добавлялись 1000 лучших биграмм для каждой меры. Отдельно следует отметить, что во всех экспериментах число тем фиксировалось равным 100.

Таблица 3

Результаты добавления 1000 лучших биграмм, упорядоченных по мере MI, в тематические модели

Корпус	Тематическая модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	PLSA	1724.2	86.1	86.1
	PLSA + биграммы	1714.1	84.2	84.2
	PLSA-SIM + биграммы	1715.4	84.1	84.1
Europarl	PLSA	1594.3	53.2	53.2
	PLSA + биграммы	1584.6	55	55
	PLSA-SIM + биграммы	1591.3	55.2	55.2
JRC-Acquiz	PLSA	812.1	67	67
	PLSA + биграммы	815.4	66.3	66.3
	PLSA-SIM + биграммы	815.6	66.4	66.4
ACL Anthology	PLSA	2134.7	74.8	74.8
	PLSA + биграммы	2138.1	75.5	75.5
	PLSA-SIM + биграммы	2144.8	75.8	75.8

Эксперименты были проведены со всеми **17** описанными выше мерами упорядочивания биграмм на всех четырех текстовых коллекциях с целью сравнить качество следующих трех алгоритмов:

1) оригинальный алгоритм PLSA;

2) алгоритм PLSA с добавленными в него как "черные ящики" 1000 лучших биграмм;

3) предложенный алгоритм PLSA-SIM с добавленными в него 1000 лучших биграмм.

В соответствии с результатами экспериментов все рассматриваемые меры распределились по двум группам.

1. В первую группу мер попали **11** ассоциативных мер: *взаимная информация (MI)*, *дополненная MI*, *нормализованная MI*, *коэффициент Сёренсена*, *симметричная условная вероятность*, *коэффициент простого соответствия*, *коэффициент Кульчинского*, *коэффициент Юла*, *коэффициент Жаккара*, *Хи-квадрат* и *отношение логарифмического правдоподобия*. При добавлении лучших биграмм по любой из данных мер в тематические модели любым из рассматриваемых способов метрики качества остаются примерно на том же самом уровне, что и в случае оригинального алгоритма PLSA. Это объясняется тем, что все эти меры упорядочивают в верх списков специфичные, нетипичные и низкочастотные биграммы, не влияющие на работу тематических моделей существенным образом. В табл. 3 представлены результаты добавления 1000 лучших биграмм, упорядоченных по мере MI (как самой широко известной в данной группе) для всех четырех текстовых коллекций. Результаты остальных мер из данной группы похожи на результаты, приведенные в табл. 3.

2. Во вторую группу попали **6** мер: *частотность (TF)*, *кубическая MI*, *настоящая MI*, *модифицированный DC*, *T-Score* и *Gravity Count*. При добавлении лучших биграмм по любой из данных мер в качестве "черных ящиков" в тематические модели перплексия ухудшается (за счет роста словаря), но улучшается согласованность тем. Следует отметить, что в работе [21] авторы в качестве меры упорядочивания биграмм рассматривали T-Score из этой группы. Таким образом, представленные

выше результаты согласуются с результатами, описанными в работе [21]. Однако при добавлении 1000 лучших биграмм по любой из мер данной группы в предложенный алгоритм PLSA-SIM значительно улучшаются все метрики качества: и перплексия, и согласованность тем. Это объясняется тем, что все эти меры упорядочивают в верх списков частотные, типичные биграммы. В табл. 4 представлены результаты добавления 1000 лучших биграмм, упорядоченных по мере TF (как самой широко известной в данной группе) для всех четырех текстовых коллекций. Результаты остальных мер из данной группы похожи на результаты, приведенные в табл. 4.

Таблица 4

Результаты добавления 1000 лучших биграмм, упорядоченных по мере TF, в тематические модели

Корпус	Тематическая модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковский	PLSA	1724.2	86.1	86.1
	PLSA + биграммы	2251.8	98.8	98.8
	PLSA-SIM + биграммы	**1450.6**	**156.5**	**102.6**
Europarl	PLSA	1594.3	53.2	53.2
	PLSA + биграммы	1993.5	57.3	57.3
	PLSA-SIM + биграммы	**1431.6**	**127.7**	**84.7**
JRC-Acquiz	PLSA	812.1	67	67
	PLSA + биграммы	1038.9	72	72
	PLSA-SIM + биграммы	**743.7**	**108.4**	**76.9**
ACL Anthology	PLSA	2134.7	74.8	74.8
	PLSA + биграммы	2619.3	73.7	73.7
	PLSA-SIM + биграммы	**1806.4**	**152.7**	**87.8**

Таким образом, при добавлении 1000 лучших биграмм, упорядоченных по любой из мер из второй группы, в предложенный алгоритм PLSA-SIM значительно улучшается качество тематических моделей по всем целевым метрикам независимо от языка и предметной области.

Помимо автоматических оценок качества тематических моделей использовались также и экспертные. Для получения экспертных оценок качества тематических моделей были приглашены двое экспертов-лингвистов. Им были выданы темы, полученные с помощью следующих трех алгоритмов:

— оригинальный алгоритм PLSA;

— алгоритм PLSA с добавлением в качестве “черных ящиков” 1000 лучших биграмм по мере TF;

— предложенный алгоритм PLSA-SIM с добавлением 1000 лучших биграмм по мере TF.

Перед экспертами была поставлена задача классификации выданных тем на 2 класса в зависимости от того, можно ли той или иной теме дать некоторое обобщенное название (класс ’+’) или нет, и тема “мусорная”, состоящая из разрозненных слов (класс ’–’). В табл. 5 представлены результаты экспертных оценок для всех текстовых коллекций, кроме архива исследовательских работ по компьютерной лингвистике ACL Anthology, поскольку для правильной разметки тем в этой коллекции требуется наличие специальных знаний в области компьютерной лингвистики.

Как видно из табл. 5, при добавлении 1000 лучших биграмм, упорядоченных по мере TF, в предложенный алгоритм PLSA-SIM количество тем, которым может быть выдано некоторое обобщенное название, увеличивается по сравнению с оригинальным алгоритмом PLSA для всех текстовых коллекций. Также следует отметить, что добавление биграмм в качестве “черных ящиков” не увеличивает число таких тем. Данный результат также подтверждает то, что предложенный алгоритм улучшает качество тематических моделей независимо от языка и предметной области.

В табл. 6 представлены первые 10 униграмм и биграмм из одной случайно выбранной темы из каждого текстового корпуса для оригинального алгоритма PLSA и предложенного алгоритма PLSA-SIM с добавлением 1000 лучших биграмм, упорядоченных по частотности (TF). В рамках одной и той же текстовой коллекции представлены темы с одинаковыми названиями, данными обоими экспертами-лингвистами.

6.2. Интеграция биграмм в тематические модели с помощью алгоритма PLSA-ITER. Предложенный итеративный алгоритм PLSA-ITER был также апробирован на описанных выше четырех

Таблица 5

Результаты экспертной оценки тем для алгоритмов PLSA и PLSA-SIM

Корпус	Тематическая модель	Количество +		Количество –	
		1 эксперт	2 эксперт	1 эксперт	2 эксперт
Банковский	PLSA	93	92	7	8
	PLSA + биграммы	92	95	8	5
	PLSA-SIM + биграммы	**96**	**97**	**4**	**3**
JRC-Acquiz	PLSA	98	90	2	10
	PLSA + биграммы	96	97	4	3
	PLSA-SIM + биграммы	**100**	**100**	**0**	**0**
Europarl	PLSA	91	99	9	1
	PLSA + биграммы	94	97	6	3
	PLSA-SIM + биграммы	**99**	**100**	**1**	**0**

Таблица 6

Первые 10 униграмм и биграмм из тем, полученных оригинальным алгоритмом PLSA и алгоритмом PLSA-SIM с добавлением 1000 лучших биграмм, упорядоченных по мере TF

Банковский корпус		Europarl		JRC-Acquiz		ACL Anthology	
PLSA	PLSA-SIM	PLSA	PLSA-SIM	PLSA	PLSA-SIM	PLSA	PLSA-SIM
Бумага	Ценная бумага	Financial	Economic crisis	Animal	Animal	Tree	Tree
Ценный	Бумага	Crisis	Financial crisis	Bovine	Bovine animal	Node	Node
Облигация	Облигация	Have	European economy	Have	Meat	Structure	Tree structure
Выпуск	Выпуск облигаций	European	Time of crisis	Slaughter	Animal health	Root	Parse tree
Рынок	Сделка	Market	Crisis	Health	Have	Label	Root node
Акция	Выпуск	Need	Current crisis	Disease	Number of animal	Figure	Decision tree
Эмитент	Сделка РЕПО	Regulation	Economic recovery	State	Bovine	Subtree	Syntactic tree
Размещение	Эмитент	System	European project	Member	Meat product	Have	Leaf node
Эмиссия	РЕПО	Supervision	Financial market	Veterinary	Test	Child	Parsing model
Обращение	Вторичный рынок	Agency	Financial	Embryo	Slaughter	Set	Tree kernel

текстовых коллекциях. В табл. 7 представлены результаты работы итеративного алгоритма PLSA-ITER после первой итерации вместе с результатами работы алгоритма PLSA-SIM с добавлением в него 1000 лучших биграмм, упорядоченных по мере TF.

Как видно из результатов, представленных в табл. 7, никакого улучшения метрик качества тематических моделей в результате работы итеративного алгоритма не происходит. Напротив, заметно их ухудшение. Это связано с тем, что на каждой итерации потенциальные кандидаты в похожие униграммы и биграммы отбираются очень тщательно, и в результате множеств похожих униграмм и биграмм образуется очень мало, недостаточно для улучшения качества. Поэтому было высказано предположение, что необходимо найти большее количество похожих униграмм и биграмм из образующихся кандидатов. Для

Таблица 7

Сравнение результатов работы итеративного алгоритма PLSA-ITER после первой итерации и алгоритма PLSA-SIM с добавлением в него 1000 лучших биграмм, упорядоченных по мере TF

Корпус	Тематическая модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковская	PLSA-SIM	1450.6	156.5	102.6
	PLSA-ITER	1499.6	138.3	97.3
Europarl	PLSA-SIM	1431.6	127.7	84.7
	PLSA-ITER	1303.6	98.6	59.3
JRC-Acquiz	PLSA-SIM	743.7	108.4	76.9
	PLSA-ITER	786.9	92.5	68.2
ACL Anthology	PLSA-SIM	1806.4	152.7	87.8
	PLSA-ITER	1949.4	119.6	77.1

Таблица 8

Примеры множеств похожих униграмм и биграмм, получающихся по разным стеммерам

Стеммер	Множество похожих униграмм и биграмм	Центральная униграмма
Snowball	Тайна, банковская тайна, тайный	Тайна
	Право, право собственности, правый, правая сторона	Право
Портер	Fish, fish agreement, fishing, fishing agreement	Fish
	Alcohol, use of alcohol, alcoholic, alcoholic product	Alcohol
Ланкастер	Budget, budget year, budgetary, budgetary year	Budget
	Culture, european culture, cultural, cultural Europe	Culture

Таблица 9

Сравнение результатов работы итеративного алгоритма PLSA-ITER со стриммерами после первой итерации и алгоритма PLSA-SIM с добавлением в него 1000 лучших биграмм, упорядоченных по мере TF

Коллекция	Тематическая модель	Перплексия	TC-PMI	TC-PMI-nSIM
Банковская	PLSA-SIM	1450.6	156.5	102.6
	PLSA-ITER + Snowball	**1265.1**	137.6	96.7
Europarl	PLSA-SIM	1431.6	127.7	84.7
	PLSA-ITER + Портер	1293.8	99.6	61.2
	PLSA-ITER + Ланкастер	**1077.7**	105	55.2
JRC-Acquiz	PLSA-SIM	743.7	108.4	76.9
	PLSA-ITER + Портер	777.7	90.8	68.2
	PLSA-ITER + Ланкастер	**736.5**	94.5	68.6
ACL Anthology	PLSA-SIM	1806.4	152.7	87.8
	PLSA-ITER + Портер	1853.7	123.6	76.2
	PLSA-ITER + Ланкастер	**1772.1**	121.3	76.5

этой цели было решено использовать стеммеры, т.е. алгоритмы, пытающиеся найти основы для заданных исходных слов.

1. Для английской части исследования были выбраны два различных стеммера:

   — широко известный стеммер Портера [44]. Данный алгоритм, применяя последовательно ряд

Таблица 10
Результаты экспертной оценки тем для алгоритмов PLSA-ITER и PLSA-SIM

Корпус	Тематическая модель	Количество +		Количество –	
		1 эксперт	2 эксперт	1 эксперт	2 эксперт
Банковский	PLSA-SIM	96	97	4	3
	PLSA-ITER	96	97	4	3
JRC-Acquiz	PLSA-SIM	100	100	0	0
	PLSA-ITER	100	100	0	0
Europarl	PLSA-SIM	99	100	1	0
	PLSA-ITER	96	99	4	1

Таблица 11
Результаты первых итераций итеративного алгоритма PLSA-ITER

Коллекция	Итерация	Перплексия	TC-PMI	TC-PMI-nSIM
Банковская	0 (PLSA)	1724.2	86.1	86.1
	1	**1265.1**	137.6	96.7
	2	1257.1	133.5	95
	3	1259.8	134.5	95.7
	4	1259.7	130.6	92.6
Europarl	0 (PLSA)	1594.3	53.2	53.2
	1	**1077.7**	105	55.2
	2	1210.8	92.1	55.2
	3	1242.9	80.1	53.2
	4	1244	84.6	52.5
JRC-Acquiz	0 (PLSA)	812.1	67	67
	1	**736.5**	94.5	68.6
	2	751.9	94.9	67
	3	749.6	99.5	67.7
	4	751.7	100.1	68.4
ACL Anthology	0 (PLSA)	2134.7	74.8	74.8
	1	**1772.1**	121.3	76.5
	2	1775.5	139.3	81
	3	1767.6	144.6	83
	4	1754.1	146.1	81.8

правил, отсекает окончания и суффиксы, основываясь на особенностях языка. Так, стеммер Портера приведет к одинаковой основе *fish* слова *fishing* и *fish*;

— более “агрессивный” стеммер Ланкастерского университета [45]. Данный алгоритм отличается от стеммера Портера тем, что применяющийся набор правил более “агрессивно” отсекает окончания и суффиксы. Так, данный стеммер приведет слова *Europe* и *European* к общей основе *europ*, в то время как стеммер Портера оставит исходные слова без изменений.

2. Для русской части исследования был выбран единственный широко известный стеммер Snowball (`http://snowball.tartarus.org/algorithms/russian/stemmer.html`). Данный алгоритм является своеобразным переводом стеммера Портера на русский язык с учетом морфологии. Так, стеммер Snowball приведет к одинаковой основе *тайн* слова *тайна* и *тайный*.

Для того чтобы учесть стеммеры в итеративном алгоритме, было модифицировано определение множеств похожих униграмм и биграмм $S = \{S_w\}$, где S_w теперь задается следующим образом:

$$S_w = \left\{ w, \bigcup_u u, \bigcup_{u,v} uv : \operatorname{stem}(u) = \operatorname{stem}(w) \quad \text{или} \quad \operatorname{stem}(v) = \operatorname{stem}(w) \right\},$$

где w и u — лемматизированные униграммы, uv — лемматизированная биграмма, $\operatorname{stem}(u)$ — основа слова u, получающаяся в результате работы того или иного стеммера. В табл. 8 приведены несколько примеров множеств похожих униграмм и биграмм по тому или иному стеммеру вместе с центральной униграммой в каждом множестве.

В табл. 9 представлены результаты работы итеративного алгоритма PLSA-ITER со стеммерами после первой итерации вместе с результатами работы алгоритма PLSA-SIM с добавлением в него 1000 лучших биграмм, упорядоченных по мере TF.

Таблица 12

Первые 10 униграмм и биграмм из тем, полученных интерактивным алгоритмом PLSA-ITER после первой итерации

Банковский корпус		Europarl		JRC-Acquiz		ACL Anthology	
Система банк	Страховой компания	European budget	Fishery agreement	Community producer	Fishing vessel	Cluster	Parse
Платежный система	Страховой	Budgetary	Fish stock	Import of product	Fishing	Clustering	Parsing
Система расчет	Страхование	Budget	Fishing	Community market	Fishery	Similarity	Parser
Система	Страховой случай	Commission budget	Fish	Community industry	Fish	Similar	Chart parsing
Банковский система	Договор страхование	Budgetary policy	Fishing agreement	Producer	Vessel	Similarity measure	Chart parser
Система платеж	Компания	Financial year	Fishing fleet	Sale of product	Fishing area	Use	Grammar
Развитие система	Страхование жизнь	Budget year	Fishery	Production	Fish stock	Vector	Use
Работа система	Страховой выплата	Have	Have	Export price	Board fishing	Method	Rule
Платежный	Страховой взнос	Budgetary year	European commission	Import price	Fishing license	Distance	Sentence
Расчет	Страховщик	European fund	Committee	Community	Board vessel	Algorithm	Edge

Как видно из результатов, представленных в табл. 9, использование стеммера Snowball для русского языка и агрессивного стеммера Ланкастерского университета в итеративном алгоритме PLSA-ITER приводит к дальнейшему улучшению качества тематических моделей по перплексии с незначительным падением уровня согласованности тем. Отдельно стоит отметить, что стеммер Портера образует немногим больше множеств похожих слов и биграмм, чем создается в алгоритме PLSA-SIM, поэтому его результаты не слишком отличаются от тех, что представлены в табл. 7.

Кроме того, были получены экспертные оценки тем, построенных с помощью первой итерации алгоритма PLSA-ITER, использующего стеммер Портера для английского языка и его модификацию для русского языка — стеммер Snowball. В табл. 10 приведены результаты экспертных оценок для всех текстовых коллекций, кроме архива исследовательских работ по компьютерной лингвистике ACL Anthology, поскольку для правильной разметки тем в этой коллекции требуется наличие специальных знаний в области компьютерной лингвистики (для сравнения приведены результаты алгоритма PLSA-SIM с добавлением 1000 лучших биграмм, упорядоченных по частотности (TF)).

Как видно из табл. 10, в результате работы алгоритма PLSA-ITER количество тем, которым можно дать некоторое обобщенное название, практически не изменяется по сравнению с алгоритмом PLSA-SIM с добавлением 1000 лучших биграмм, упорядоченных по частотности.

В табл. 11 представлены результаты первых итераций итеративного алгоритма PLSA-ITER со стеммерами Snowball и Ланкастерского университета для всех текстовых коллекций вместе с результатами оригинального алгоритма PLSA (в таблице обозначен как нулевая итерация).

Как видно из результатов, представленных в табл. 11, после первой итерации наблюдается существенное улучшение качества полученных тем. Стоит заметить, что на следующих итерациях результаты начинают колебаться вокруг примерно тех же самых уровней перплексии и согласованности тем.

В табл. 12 представлены первые 10 слов и биграмм из двух случайно выбранных тем из каждого текстового корпуса для предложенного алгоритма PLSA-ITER со стеммерами Ланкастерского университета и Snowball после первой итерации.

7. Заключение. В настоящей статье представлены эксперименты по добавлению биграмм и сходства между ними и униграммными компонентами в тематические модели. Вначале предлагается новый алгоритм PLSA-SIM, добавляющий похожие униграммы и биграммы в тематические модели и учитывающий взаимосвязь между биграммами и униграммными компонентами. Эксперименты, проведенные на английских частях многоязычных корпусов параллельных текстов Europarl и JRC-Acquiz, архиве исследований по компьютерной лингвистике ACL Anthology и подборке русскоязычных банковских статей, выделили две группы мер, упорядочивающих биграммы, встретившиеся в коллекции. Добавление верхних частей списков биграмм, упорядоченных по первой группе мер, в тематическую модель любым способом не меняет качество тематических моделей. Однако добавление верхних частей списков биграмм, упорядоченных по второй группе мер, в алгоритм PLSA-SIM приводит к существенному улучшению качества тематических моделей по всем рассматриваемым метрикам. Кроме того, предлагается еще один новый итеративный алгоритм, позволяющий добавлять наиболее подходящие биграммы и похожие слова. Проведенные эксперименты показывают, что использование стеммеров в предложенном алгоритме позволяет еще более улучшить качество тематических моделей по перплексии. Предложенные модификации также были применены и к алгоритму LDA, и результаты оказались аналогичными.

Работа частично поддержана грантом РФФИ 14–07–00383.

СПИСОК ЛИТЕРАТУРЫ

1. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. **3**. 993–1022.
2. *Wei X., Croft W.B.* LDA-based document models for ad-hoc retrieval // Proceedings of the 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006. 178–185.
3. *Boyd-Graber J.L., Blei D.M., Zhu X.* A topic model for word sense disambiguation // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL Press, 2007. 1024–1033.
4. *Wang D., Zhu S., Li T., Gong Y.* Multi-document summarization using sentence-based topic models // Proceedings of the ACL-IJCNLP Conference Short Papers. Stroudsburg: ACL Press, 2009. 297–300.
5. *Eidelman V., Boyd-Graber J., Resnik P.* Topic models for dynamic translation model adaptation // Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics. Short Papers. Vol. 2. Stroudsburg: ACL Press, 2012. 115–119.
6. *Zhou S., Li K., Liu Y.* Text categorization based on topic model // International Journal of Computational Intelligence Systems. 2009. **2**, N 4. 398–409.
7. *Bolelli L., Ertekin Ş., Giles C.L.* Topic and trend detection in text collections using latent Dirichlet allocation // Lecture Notes in Computer Science. Vol. 5478. Heidelberg: Springer, 2009. 776–780.
8. *Hyunh T., Fritz M., Schiele B.* Discovery of activity patterns using topic models // Proceedings of the 10th International Conference on Ubiquitous Computing. New York: ACM Press, 2008. 10–19.
9. *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999. 50–57.
10. *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. 2010. **4**, N 2. 280–301.
11. *Wallach H.M.* Topic modeling: beyond bag-of-words // Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006. 977–984.
12. *Griffiths T.L., Steyvers M., Tenenbaum J.B.* Topics in semantic representation // Psychological Review. 2007. **144**, N 2. 211–244.
13. *Wang X., McCallum A., Wei X.* Topical n-grams: phrase and topic discovery, with an application to information retrieval // Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. Washington: IEEE Press, 2007. 697–702.
14. *He Q., Chang K., Lim E., Banerjee A.* Keep it simple with time: a reexamination of probabilistic topic detection models // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010. **32**, N 10. 1795–1808.

15. *Salton G.* Automatic text processing: the transformation, analysis, and retrieval of information by computer. Boston: Addison-Wesley, 1989.
16. *MacQueen J.* Some methods for classification and analysis of multivariate observations // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967. 281–297.
17. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society. Series B. 1977. **39**, N 1. 1–38.
18. *Asuncion A., Welling M., Smyth P., Teh Y.W.* On smoothing and inference for topic models // Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Arlington: AUAI Press, 2009. 27–34.
19. *Hu W., Shimizu N., Nakagawa H., Sheng H.* Modeling Chinese documents with topical word-character models // Proceedings of the 22nd International Conference on Computational Linguistics. Stroudsburg: ACL Press, 2008. 345–352.
20. *Johnson M.* PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names // Proceedings of the 48th Annual Meeting of the ACL. Stroudsburg: ACL Press, 2010. 1148–1157.
21. *Lau J.H., Baldwin T., Newman D.* On collocations and topic models // ACM Transactions on Speech and Language Processing. 2013. **10**, N 3. 1–14.
22. *Newman D., Lau J.H., Grieser K., Baldwin T.* Automatic evaluation of topic coherence // Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies. Stroudsburg: ACL Press, 2010. 100–108.
23. *Andrzejewski D., Zhu X., Craven M.* Incorporating domain knowledge into topic modeling via Dirichlet forest priors // Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009. 25–32.
24. *Liu B.* Sentiment analysis and opinion mining. San Rafael: Morgan & Claypool Publishers, 2012.
25. *Zhai Z., Liu B., Xu H., Jia P.* Grouping product features using semi-supervised learning with soft-constraints // Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg: ACL Press, 2010. 1272–1280.
26. *Воронцов К.В., Потапенко А.А.* Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. **1**, № 6. 657–686.
27. *Chang J., Boyd-Graber J., Wang C., Gerrich S., Blei D.* Reading tea leaves: how human interpret topic models // Proceedings of the 24th Annual Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2009. 288–296.
28. *Mimno D., Wallach H.M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proc. Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2011. 262–272.
29. *Stevens K., Kegelmeyer P., Andrzejewski D., Butter D.* Exploring topic coherence over many models and many topics // Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL Press, 2012. 952–961.
30. *Andrzejewski D., Butter D.* Latent topic feedback for information retrieval // Proceedings of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011. 600–608.
31. *Church K.W., Hanks P.* Word association norms, mutual information, and lexicography // Computational Linguistics. 1990. Vol. 16, N 1. 22–29.
32. *Zhang W., Yoshida T., Ho T.B., Tang X.* Augmented mutual information for multi-word extraction // International Journal of Innovative Computing, Information and Control. 2009. **5**, N 2. 543–554.
33. *Bouma G.* Normalized (pointwise) mutual information in collocation extraction // Proceedings of the Biennial GSCL Conference. Tübingen: Gunter Narr Verlag, 2009. 31–40.
34. *Deane P.A.* A nonparametric method for extraction of candidate phrasal terms // Proceedings of the 43rd Annual Meeting of the ACL. Stroudsburg: ACL Press, 2005. 605–613.
35. *Daille B.* Combined approach for terminology extraction: lexical statistics and linguistic filtering. PhD Thesis. Paris: University of Paris, 1995.
36. *Silva J.F., Lopes G.P.* A local maxima method and a fair dispersion normalization for extracting multiword units from corpora // Proceedings of the 6th Meeting on the Mathematics of Language. Stroudsburg: ACL Press, 1999. 369–381.
37. *Smadja F., McKeown K.R., Hatzivassiloglou V.* Translating collocations for bilingual lexicons: a statistical approach // Computational Linguistics. 1996. **22**, N 1. 1–38.
38. *Kitamura M., Matsumoto Y.* Automatic extraction of word sequence correspondences in parallel corpora // Proceedings of the 4th Annual Workshop on Very Large Corpora. Stroudsburg: ACL Press, 1996. 79–87.
39. *Daudaravičius V., Marcinkevičienė R.* Gravity counts for the boundaries of collocations // Int. J. Corpus Linguistics. 2004. **9**, N 2. 321–348.
40. *Kulczińsky S.* Zespoly róslin w Pieninach (Die Pflanzenassociationen der Pienenen) // Bulletin International de L'Acad'emie Polonaise des Sciences et des Letters. Classe des Sciences Mathematiques et Naturelles. Serie B. Suppl'ement II. 1927. N 2. 57–203.
41. *Jaccard P.* Distribution de la flore alpine dans le bassin des drances et dans quelques régions voisines // Bull. Soc. Vaudoise Sci. Natur. 1901. **37**. 241–272.

42. *Gale W.A., Church K.W.* A Program for aligning sentences in bilingual corpora // Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 1991. 177–184.
43. *Dunning T.* Accurate methods for the statistics of surprise and coincidence // Computational Linguistics. 1993. **19**, N 1. 61–74.
44. *Porter M.F.* An algorithm for suffix stripping // Program. 1980. **14**, N 3. 130–137.
45. *Paice C.D.* Another stemmer // ACM SIGIR Forum. 1990. **24**, N 3. 56–61.

Поступила в редакцию
12.03.2015

Topic Models: Adding Bigrams and Taking Account of the Similarity between Unigrams and Bigrams

M. A. Nokel [1] **and N. V. Loukachevitch** [2]

[1] *Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics; Leninskie Gory, Moscow, 119991, Russia; Graduate Student, e-mail: mnokel@gmail.com*

[2] *Research Computing Center, Lomonosov Moscow State University; Leninskie Gory, Moscow, 119992, Russia; Ph.D., Leading Scientist, e-mail: louk_nat@mail.ru*

Received March 12, 2015

Abstract: The results of experimental study of adding bigrams and taking account of the similarity between them and unigrams are discussed. A novel PLSA-SIM algorithm based on a modification of the original PLSA (Probabilistic Latent Semantic Analysis) algorithm is proposed. The proposed algorithm incorporates bigrams and takes into account the similarity between them and unigram components. Various word association measures are analyzed to integrate top-ranked bigrams into topic models. As target text collections, articles from various Russian electronic banking magazines, English parts of parallel corpora Europarl and JRC-Acquiz, and the English digital archive of research papers in computational linguistics (ACL Anthology) are chosen. The computational experiments show that there exists a subgroup of tested measures that produce top-ranked bigrams in such a way that their inclusion into the PLSA-SIM algorithm significantly improves the quality of topic models for all collections. A novel unsupervised iterative algorithm named PLSA-ITER is also proposed for adding the most relevant bigrams. The computational experiments show a further improvement in the quality of topic models compared to the PLSA algorithm.

Keywords: topic models, PLSA (Probabilistic Latent Semantic Analysis), word association measures, bigrams, topic coherence, perplexity.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res. **3**, 993–1022 (2003).

2. X. Wei and W. B. Croft, "LDA-Based Document Models for Ad-hoc Retrieval," in *Proc. 29th Annual Int. ACM-SIGIR Conf. on Research and Development in Information Retrieval, Seattle, USA, August 6–10, 2006* (ACM Press, New York, 2006), pp. 178–185.

3. J. L. Boyd-Graber, D. M. Blei, and X. Zhu, "A Topic Model for Word Sense Disambiguation," in *Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, June 28–30, 2007* (ACL Press, Stroudsburg, 2007), pp. 1024–1033.

4. D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-Document Summarization Using Sentence-Based Topic Models," in *Proc. ACL-IJCNLP Conf. Short Papers, Singapore, Singapore, August 2–7, 2009* (ACL Press, Stroudsburg, 2009), pp. 297–300.

5. V. Eidelman, J. Boyd-Graber, and P. Resnik, "Topic Models for Dynamic Translation Model Adaptation," in *Proc. 50th Annual Meeting of the Association of Computational Linguistics, Stroudsburg, USA, Short Papers, July 8–14, 2012* (ACL Press, Stroudsburg, 2012), Vol. 2, pp. 115–119.

6. S. Zhou, K. Li, and Y. Liu, "Text Categorization Based on Topic Model," Int. J. Comput. Intell. Syst. **2** (4), 398–409 (2009).

7. L. Bolelli, Ş. Ertekin, and C. L. Giles, "Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation," in *Lecture Notes in Computer Science* (Springer, Heidelberg, 2009), Vol. 5478, pp. 776–780.

8. T. Hyunh, M. Fritz, and B. Schiele, "Discovery of Activity Patterns Using Topic Models," in *Proc. 10th Int. Conf. on Ubiquitous Computing, Seoul, South Korea, September 21–24, 2008* (ACM Press, New York, 2008), pp. 10–19.

9. T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. of the 22nd Annual Int. SIGIR Conf. on Research and Development in Information Retrieval, Berkley, USA, August 15–19, 1999* (ACM Press, New York, 1999), pp. 50–57.

10. A. Daud, J. Li, L. Zhou, and F. Muhammad, "Knowledge Discovery through Directed Probabilistic Topic Models: A Survey," Front. Comput. Sci. China **4** (2), 280–301 (2010).

11. H. M. Wallach, "Topic Modeling: Beyond Bag-of-Words," in *Proc. 23rd Int. Conf. on Machine Learning, Pitsburg, USA, June 25–29, 2006* (ACM Press, New York, 2006), pp. 977–984.

12. T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in Semantic Representation," Psychol. Rev. **144** (2), 211–244 (2007).

13. X. Wang, A. McCallum, and X. Wei, "Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval," in *Proc. 7th IEEE Int. Conf. on Data Mining, Las Vegas, USA, October 28–31, 2007* (IEEE Press, Washington, DC, 2007), pp. 697–702.

14. Q. He, K. Chang, E. Lim, and A. Banerjee, "Keep It Simple with Time: A Reexamination of Probabilistic Topic Detection Models," IEEE Trans. Pattern Anal. Mach. Intell. **32** (10), 1795–1808 (2010).

15. G. Salton, *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley, Boston, 1989).

16. J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, June 21–July 18, 1965 and December 27, 1965–January 7, 1966* (Univ. California Press, Berkeley, 1967), pp. 281–297.

17. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Roy. Stat. Soc., Series B Stat. Methodol. **39** (1), 1–38 (1977).

18. A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On Smoothing and Inference for Topic Models," in *Proc. 25th Conf. on Uncertainty in Artificial Intelligence, Montreal, Canada, June 18–21, 2009* (AUAI Press, Arlington, 2009), pp. 27–34.

19. W. Hu, N. Shimizu, H. Nakagawa, and H. Sheng, "Modeling Chinese Documents with Topical Word-Character Models," in *Proc. 22nd Int. Conf. on Computational Linguistics, Manchester, UK, August 18–22, 2008* (ACL Press, Stroudsburg, 2008), pp. 345–352.

20. M. Johnson, "PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names," in *Proc. 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, July 11–16, 2010* (ACL Press, Stroudsburg, 2010), pp. 1148–1157.

21. J. H. Lau, T. Baldwin, and D. Newman, "On Collocations and Topic Models," ACM Trans. Speech Lang. Process. **10** (3), 1–14 (2013).

22. D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," in *Proc. 11th Annual Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, Los Angeles, USA, June 1–6, 2010* (ACL Press, Stroudsburg, 2010), pp. 100–108.

23. D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors," in *Proc. 26th Annual Int. Conf. on Machine Learning, Montreal, Canada, June 14–18, 2009* (ACM Press, New York, 2009), pp. 25–32.

24. B. Liu, *Sentiment Analysis and Opinion Mining* (Morgan & Claypool, San Rafael, 2012).

25. Z. Zhai, B. Liu, H. Xu, and P. Jia, "Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints," in *Proc. 23rd Int. Conf. on Computational Linguistics, Beijing, China, August 23–27, 2010* (ACL Press, Stroudsburg, 2010), pp. 1272–1280.

26. K. V. Vorontsov and A. A. Potapenko, "EM-like Algorithms for Probabilistic Topic Modeling," Mashin. Obuchenie Analiz Dannykh **1** (6), 657–686 (2013).

27. J. Chang, J. Boyd-Graber, C. Wang, et al., "Reading Tea Leaves: How Human Interpret Topic Models," in *Proc. 24th Annual Conf. on Neural Information Processing Systems, Vancouver, Canada, December 6–9, 2010* (Curran Associates, Red Hook, 2010), pp. 288–296.

28. D. Mimno, H. M. Wallach, E. Talley, et al., "Optimizing Semantic Coherence in Topic Models," in *Proc. Conf. on Empirical Methods in Natural Language Processing, Edinburgh, UK, July 27–29, 2011* (ACL Press, Stroudsburg, 2011), pp. 262–272.

29. K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Butter, "Exploring Topic Coherence over Many Models and Many Topics," in *Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju, Korea, July 12–14, 2012* (ACL Press, Stroudsburg, 2012), pp. 952–961.

30. D. Andrzejewski and D. Butter, "Latent Topic Feedback for Information Retrieval," in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Diego, USA, August 21–24, 2011* (ACM Press, New York, 2011), pp. 600–608.

31. K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," Comput. Linguist. **16** (1), 22–29 (1990).

32. W. Zhang, T. Yoshida, T.B. Ho, and X. Tang, "Augmented Mutual Information for Multi-Word Extraction," Int. J. Innov. Comput. Inform. Contr. **5** (2), 543–554 (2009).

33. G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," in *Proc. Biennial GSCL Conf., Potsdam, Germany, September 30–October 2, 2009* (Gunter Narr Verlag, Tübingen, 2009), pp. 31–40.

34. P. A. Deane, "A Nonparametric Method for Extraction of Candidate Phrasal Terms," in *Proc. 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, USA, June 25–30, 2005* (ACL Press, Stroudsburg, 2005), pp. 605–613.

35. B. Daille, *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering*, PhD Thesis (Univ. of Paris, Paris, 1995).

36. J. F. Silva and G. P. Lopes, "A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units from Corpora," in *Proc. 6th Meeting on the Mathematics of Language, Florida, USA, July 23–25, 1999* (ACL Press, Stroudsburg, 1999), pp. 369–381.

37. F. Smadja, K. R. McKeown, and V. Hatzivassiloglou, "Translating Collocations for Bilingual Lexicons: A Statistical Approach," Comput. Linguist. **22** (1), 1–38 (1996).

38. M. Kitamura and Y. Matsumoto, "Automatic Extraction of Word Sequence Correspondences in Parallel Corpora," in *Proc. 4th Annual Workshop on Very Large Corpora, Copenhagen, Denmark, August 4, 1996* (ACL Press, Stroudsburg, 1996), pp. 79–87.

39. V. Daudaravičius and R. Marcinkevičiené, "Gravity Counts for the Boundaries of Collocations," Int. J. Corpus Linguist. **9** (2), 321–348 (2004).

40. S. Kulczińsky, "Zespoly róslin w Pieninach (Die Pflanzenassociationen der Pienenen)," Bull. Int. de L'Acad'emie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles, Serie B, Suppl. II, No. 2, 57–203 (1927).

41. P. Jaccard, "Distribution de la Flore Alpine dans le Bassin des Drances et dans Quelques Régions Voisines," Bull. Soc. Vaudoise Sci. Natur. **37**, 241–272 (1901).

42. W. A. Gale and K. W. Church, "A Program for Aligning Sentences in Bilingual Corpora," in *Proc. 29th Annual Meeting of the Association for Computational Linguistics, Berkley, USA, June 18–21, 1991* (ACL Press, Stroudsburg, 1991), pp. 177–184.

43. T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Comput. Linguist. **19** (1), 61–74 (1993).

44. M. F. Porter, "An algorithm for suffix stripping," Program **14** (3), 130–137 (1980).

45. C. D. Paice, "Another Stemmer," ACM SIGIR Forum **24** (3), 56–61 (1990).