УДК 004.852:519.766.4 doi 10.26089/NumMet.v16r104

РЕГУЛЯРИЗАЦИЯ МНОГОЯЗЫЧНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

М. А. Дударенко[1]

Предлагается многоязычная вероятностная тематическая модель, одновременно учитывающая двуязычный словарь и связи между документами параллельной или сравнимой коллекции. Для комбинирования этих двух видов информации применяется аддитивная регуляризация тематических моделей (ARTM). Предлагаются два способа использования двуязычного словаря: первый учитывает только сам факт связи между словами–переводами, во втором настраиваются вероятности переводов в каждой теме. Качество многоязычных моделей измеряется на задаче кросс-язычного поиска, когда запросом является документ на одном языке, а поиск производится среди документов другого языка. Показано, что комбинированный учет слов–переводов из двуязычного словаря и связанных документов улучшает качество кросс-язычного поиска по сравнению с моделями, использующими только один тип информации. Сравнение разных методов включения в модель двуязычных словарей показывает, что оценивание вероятностей переводов не только улучшает качество модели, но и позволяет находить тематический контекст для пар “слово–перевод”.

Ключевые слова: многоязычная тематическая модель, вероятностная тематическая модель, параллельная коллекция, сравнимая коллекция, двуязычный словарь, регуляризация, кросс-язычный поиск.

1. Введение. Вероятностные тематические модели (ВТМ) применяются для выявления латентных тем в коллекциях текстовых документов на естественных языках. Такие модели представляют каждый документ вероятностным распределением на множестве тем, а каждую тему — вероятностным распределением на множестве слов из словаря коллекции.

ВТМ могут быть использованы для классификации и кластеризации документов, определения тематики документов, поиска документов схожей тематики, рекомендации документов и др. Данная задача актуальна и для многоязычных коллекций. Построение многоязычных тематических моделей дает возможность сравнивать документы разных языков без построения их переводов с одного языка на другой, так как тематическое пространство является инвариантным относительно языка.

Тематические модели PLSA (Probabilistic Latent Semantic Analysis) [1] и LDA (Latent Dirichlet Allocation) [2], считающиеся де-факто стандартными, предполагают, что все документы написаны на одном языке. В одноязычных моделях слова объединяются в темы на основании их взаимного употребления. В случае многоязычных коллекций, даже если слова разных языков имеют схожее значение, они едва ли появятся в одном контексте, так как каждый документ написан на своем языке. Поэтому стандартные тематические модели не могут формировать многоязычные темы на основе статистики совместного употребления слов. Так, в двуязычной коллекции документов одноязычная тематическая модель находит похожие темы, но в каждой из них содержатся слова только одного языка [3, 4]. Слияние параллельных или сравнимых документов в один объединенный документ и дальнейшее применение одноязычной модели тоже дает неудовлетворительные результаты. Эксперименты показывают, что в полученных темах среди наиболее вероятных слов преобладают слова одного языка [5]. Для связывания таких тем необходимо в явном виде вводить в модель дополнительную информацию о соответствии между языками.

В многоязычной коллекции каждый язык имеет свой словарь, поэтому каждая тема представляется набором вероятностных распределений над словарями каждого из языков. Кроме того, представления темы в разных языках должны быть согласованы. Для связывания различных языков модели ML–LDA (MultiLingual LDA) [6], PLTM (PolyLingual Topic Model) [7] и BiLDA (Bilingual LDA) [8] используют коллекции документов, выровненные по словам, по предложениям или по документам. Кроме того, могут использоваться двуязычные словари или структурированные базы знаний, например в многоязычных моделях MuTo (Multilingual Topic) [3], JointLDA [4], PCLSA (Probabilistic Cross-Lingual Latent Semantic Analysis) [9] и MLSLDA (Multilingual Supervised LDA) [10].

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 119992, Москва; аспирант, e-mail: m.dudarenko@gmail.com

© Научно-исследовательский вычислительный центр МГУ им. М. В. Ломоносова

Известные многоязычные BTM используют либо параллельные или сравнимые коллекции, либо двуязычные словари, не объединяя два эти вида информации в рамках одной модели. Возможно, это объясняется техническими трудностями комбинирования моделей в байесовском подходе, основанном на использовании распределений Дирихле. В настоящей статье для комбинирования многоязычных тематических моделей используется небайесовский многокритериальный подход — аддитивная регуляризация тематических моделей, ARTM (Additive Regularization of Topic Models) [11–13]. В отличие от других подходов к регуляризации тематических моделей [14–17] подход ARTM позволяет комбинировать в EM-алгоритме (Expectation-Maximization algorithm) любое число произвольных регуляризаторов при построении многоцелевых тематических моделей.

Цель данной статьи — показать, что совместный учет параллельности коллекции и двуязычных словарей улучшает качество многоязычной тематической модели. Сравнивается влияние различных источников многоязычной информации по отдельности и совместно на качество кросс-язычного поиска. Предложенный метод сравнивается с известными реализациями многоязычных тематических моделей и с методом пословного перевода документов, основанным на использовании двуязычного словаря.

2. Постановка задачи.

2.1. Одноязычное тематическое моделирование. Пусть D — коллекция текстовых документов, W — словарь всех употребляемых в них слов (или словосочетаний). Предполагается, что существует множество T скрытых переменных–тем и каждое слово $w \in W$ в документе $d \in D$ связано с некоторой темой $t \in T$. Коллекция документов рассматривается как множество троек (d, w, t), которые генерируются случайно и независимо из дискретного распределения $p(d, w, t)$, определенного на множестве $D \times W \times T$. Это предположение называют также гипотезой “мешка слов”.

Вероятностная тематическая модель описывает условные распределения слов в документах $p(w|d)$, выражая их через условные распределения слов в темах $\phi_{wt} \equiv p(w|t)$ и тем в документах $\theta_{td} \equiv p(t|d)$:

$$p(w|d) = \sum_{t\in T} p(w|t)\, p(t|d) = \sum_{t\in T} \phi_{wt}\, \theta_{td}. \tag{1}$$

Выражение (1) следует из формулы полной вероятности и гипотезы условной независимости слов в темах от документов: $p(w|d, t) = p(w|t)$.

Задача тематического моделирования заключается в оценивании параметров модели ϕ_{wt} и θ_{td} по наблюдаемым частотам слов $\widehat{p}(w|d) = \dfrac{n_{wd}}{n_d}$, где n_{wd} — число вхождений слова w в документ d.

Для поиска параметров модели ϕ_{wt} и θ_{td} максимизируется логарифм правдоподобия [1]

$$\mathcal{L}(\Phi, \Theta) = \sum_{d\in D} \sum_{w\in d} n_{wd} \log \sum_{t\in T} \phi_{wt}\, \theta_{td} \to \max_{\Phi,\Theta}$$

при ограничениях нормировки и неотрицательности распределений:

$$\sum_{w\in W} \phi_{wt} = 1,\ \phi_{wt} \geqslant 0\ \forall\, t \in T; \quad \sum_{t\in T} \theta_{td} = 1,\ \theta_{td} \geqslant 0\ \forall\, d \in D. \tag{2}$$

Рассматриваемая задача может быть решена с помощью EM-алгоритма [1], в котором на E-шаге оцениваются вероятности скрытых тем $t \in T$ для каждой пары (d, w) документ–слово, а на M-шаге вычисляются новые значения тематических переменных ϕ_{wt} и θ_{td}:

$$\text{E-шаг:} \qquad p(t|d, w) = \frac{\phi_{wt}\, \theta_{td}}{\sum\limits_{s\in T} \phi_{ws}\, \theta_{sd}};$$

$$\text{M-шаг:} \qquad \begin{cases} \phi_{wt} = \dfrac{n_{wt}}{n_t}, & n_t = \sum\limits_{w\in W} n_{wt}, \qquad n_{wt} = \sum\limits_{d\in D} n_{wd}\, p(t|d, w); \\ \theta_{td} = \dfrac{n_{td}}{n_d}, & n_d = \sum\limits_{t\in T} n_{td}, \qquad n_{td} = \sum\limits_{w\in d} n_{wd}\, p(t|d, w). \end{cases} \tag{3}$$

Для упрощения записи будем опускать операцию нормировки и использовать знак пропорциональности $\propto$: $\phi_{wt} \propto n_{wt}$; $\theta_{td} \propto n_{td}$.

2.2. Многоязычное тематическое моделирование. Пусть теперь коллекция D содержит документы на разных языках, L — множество различных языков, представленных в коллекции, $\ell_d \in L$ —

язык, на котором написан документ d. В этом случае полная коллекция документов разбивается на $|L|$ непересекающихся одноязычных коллекций $D_\ell = \{d \in D \,|\, \ell_d = \ell\}$, каждой из которых соответствует свой одноязычный словарь W_ℓ, $\ell \in L$.

Как и в случае одного языка, предполагается существование множества T скрытых тем и вероятностного пространства $D \times W \times T \times L$ с дискретным распределением $p(d, w, t, \ell)$. В многоязычных тематических моделях обычно предполагается, что множество тем — общее для всех языков и для каждой темы $t \in T$ в каждом языке $\ell \in L$ существует свое дискретное вероятностное распределение $p(w|t, \ell)$ на множестве слов $w \in W_\ell$. Предполагается также, что для каждой темы распределения из разных языков согласованы, т.е. наиболее вероятными являются похожие по смыслу слова. Таким образом, каждая тема имеет свое представление в каждом языке. Распределения $p(t|d, \ell)$ на общем для всех языков множестве тем $t \in T$ определены для документов $d \in D_\ell$ для любого языка коллекции $\ell \in L$, что возможно благодаря многоязычному представлению каждой темы.

Задачей многоязычного тематического моделирования является построение семейств распределений $p(w|t, \ell) \equiv \phi^\ell_{wt}$ и $p(t|d, \ell) \equiv \theta^\ell_{td}$. Оценка вероятностей $\widehat{p}(w|d, \ell) = \dfrac{n^\ell_{wd}}{n^\ell_d}$ производится по словам $w \in W_\ell$ из словаря на языке $\ell \in L$ документа $d \in D_\ell$.

Для построения многоязычной тематической модели также ставится задача максимизации логарифма правдоподобия

$$\mathcal{L}(\Phi, \Theta) = \sum_{\ell \in L} \sum_{d \in D_\ell} \sum_{w \in d} n^\ell_{wd} \log \sum_{t \in T} \phi^\ell_{wt}\, \theta^\ell_{td} \to \max_{\Phi, \Theta} \tag{4}$$

при ограничениях вида (2) с поправкой на несколько языков:

$$\sum_{w \in W_\ell} \phi^\ell_{wt} = 1,\ \phi^\ell_{wt} \geqslant 0\ \forall\, t \in T,\ \forall\, \ell \in L; \quad \sum_{t \in T} \theta^\ell_{td} = 1,\ \theta^\ell_{td} \geqslant 0\ \forall\, d \in D_\ell,\ \forall \ell \in L. \tag{5}$$

Формулы E-шага и M-шага EM-алгоритма (3) тоже изменяются с поправкой на язык:

$$\begin{aligned} &\text{E-шаг:} \quad p(t|d, w, \ell) \propto \phi^\ell_{wt}\, \theta^\ell_{td}; \\ &\text{M-шаг:} \quad \phi^\ell_{wt} \propto n^\ell_{wt} = \sum_{d \in D_\ell} n^\ell_{wd}\, p(t|d, w, \ell); \quad \theta^\ell_{td} \propto n^\ell_{td} = \sum_{w \in d} n^\ell_{wd}\, p(t|d, w, \ell). \end{aligned} \tag{6}$$

Для согласования тематических распределений ϕ^ℓ_{wt} требуются дополнительные ограничения, учитывающие многоязычную информацию, связывающую разные языки друг с другом. В нашей работе для этого предлагается использовать подход аддитивной регуляризации тематических моделей ARTM [11–13].

3. Регуляризация тематических моделей. Аддитивная регуляризация тематических моделей ARTM предполагает, что заданы n критериев $\mathcal{R}_i(\Phi, \Theta)$, называемых регуляризаторами, каждый из которых нужно максимизировать и которые суммируются с логарифмом правдоподобия с неотрицательными коэффициентами τ_i: $\mathcal{L}(\Phi, \Theta) + \mathcal{R}(\Phi, \Theta) \to \max\limits_{\Phi, \Theta}$, $\mathcal{R}(\Phi, \Theta) = \sum\limits_{i=1}^{n} \tau_i\, \mathcal{R}_i(\Phi, \Theta)$.

Решение данной задачи [11] приводит к модифицированным формулам M-шага (6):

$$\phi^\ell_{wt} \propto \left(n^\ell_{wt} + \phi^\ell_{wt} \frac{\partial \mathcal{R}(\Phi, \Theta)}{\partial \phi^\ell_{wt}} \right)_+; \quad \theta^\ell_{td} \propto \left(n^\ell_{td} + \theta^\ell_{td} \frac{\partial \mathcal{R}(\Phi, \Theta)}{\partial \theta^\ell_{td}} \right)_+. \tag{7}$$

Здесь $(z)_+ = \max(z, 0)$ — положительная срезка, необходимая для выполнения ограничений неотрицательности. В случае многоязычных моделей функционалы $\mathcal{R}_i(\Phi, \Theta)$ формализуют связи между языками и учитывают различные источники многоязычной информации, при этом получение формул M-шага сводится к дифференцированию регуляризаторов по параметрам модели.

4. Учет параллельных и сравнимых коллекций. Первые многоязычные BTM строились с использованием параллельных или сравнимых коллекций документов.

Параллельная коллекция состоит из групп документов, являющихся переводами друг друга на нескольких языках. Примером такой коллекции является EuroParl [18] — корпус протоколов заседаний Европейского парламента, включающий в себя версии на 21 европейском языке. Сравнимая коллекция состоит из групп документов, которые являются скорее пересказами, чем переводами, согласованы по тематике, но могут существенно различаться даже по длине. Примерами таких коллекций являются Википедия или многоязычные издания журналов и газет. Далее группы параллельных или сравнимых документов

будем называть связками, или связанными документами. Вообще говоря, связка не обязана содержать документы на всех языках, некоторые из них могут отсутствовать.

Преимуществом многоязычных тематических моделей, основывающихся на использовании сравнимых и параллельных коллекций, является независимость от дополнительных лингвистических ресурсов: чтобы построить тематическую модель, достаточно иметь только подходящую коллекцию документов. Недостатком является сложность формирования больших параллельных коллекций.

Одними из первых многоязычных тематических моделей были ML–LDA (MultiLingual LDA) [6], PLTM (PolyLingual Topic Model) [7] и BiLDA (Bilingual LDA) [8]. Несмотря на различные названия, по сути это одно и то же естественное обобщение одноязычной тематической модели LDA.

Будем обозначать через $d = \{d_1, \ldots, d_{|L|}\}$ многоязычный *метадокумент*, где $d_\ell \in D_\ell$ — *представление* метадокумента d в языке $\ell \in L$. Назовем *тематическим профилем* объекта x распределение $p(t|x)$. Через θ^ℓ_{td} будем обозначать тематический профиль $p(t|d,\ell)$ документа d в языке ℓ.

Предположим, что тематические профили связанных документов равны: $\theta^\ell_{td} = \theta_{td},\ \ell \in L$. Для сравнимых документов данное допущение делается ради упрощения модели, так как их тематические профили все же могут различаться. Для параллельных коллекций это предположение верно, так как параллельные документы являются переводами.

Вероятностная тематическая модель документа d_ℓ обобщает одноязычную модель (1):

$$p(w|d,\ell) = \sum_{t\in T} p(w|t,\ell)\, p(t|d) = \sum_{t\in T} \phi^\ell_{wt}\, \theta_{td}.$$

Максимизация логарифма правдоподобия

$$\mathcal{L}(\Phi,\Theta) = \sum_{d\in D} \sum_{\ell\in L} \sum_{w\in d_\ell} n^\ell_{wd} \log \sum_{t\in T} \phi^\ell_{wt}\, \theta_{td} \to \max_{\Phi,\Theta} \tag{8}$$

с ограничениями нормировки дает следующие выражения М-шага ЕМ-алгоритма:

$$\phi^\ell_{wt} \propto n^\ell_{wt}; \quad \theta_{td} \propto n_{td} = \sum_{\ell\in L} n^\ell_{td}. \tag{9}$$

Из последнего выражения для θ_{td} в (9) видно, что от обычной формулы оно отличается суммированием по всем связанным документам.

Данная модель может использоваться также тогда, когда только часть многоязычной коллекции содержит связанные документы. В этом случае максимизируется сумма логарифмов правдоподобия (8) для документов–переводов или (4) для тех документов, переводы которых не известны. Этот метод обучения многоязычных тематических моделей будем называть ML–P (MultiLingual Parallel).

5. Учет словарей. При использовании параллельных или сравнимых коллекций для согласования тематических распределений накладываются ограничения на тематические профили θ_{td} связанных документов. Двуязычные словари, наоборот, накладывают ограничения на тематические матрицы ϕ^ℓ_{wt}, связывая слова–переводы. Методы обучения многоязычных тематических моделей, использующие внешние двуязычные словари или структурированные базы знаний, такие как WordNet [19], EuroWordNet [20], BabelNet [21], DBpedia [22] и MENTA [23], отличаются бóльшим разнообразием, поскольку существует много способов учитывать связи между словами и неполноту словарей, допуская отсутствие в словаре любого слова или его перевода.

Модели MuTo (Multilingual Topic) [3] и JointLDA [4] основаны на модели LDA. Для получения многоязычных распределений вводится понятие *концепта* — пары слов, являющихся переводами. Тематические распределения в данных моделях строятся на концептах, а не на словах разных языков. Таким образом обеспечивается согласованность многоязычных распределений, однако число языков в коллекции ограничивается двумя. Обобщение на большее число языков представляется затруднительным. Модель PCLSA (Probabilistic Cross-Lingual Latent Semantic Analysis) [9] является расширением модели PLSA с помощью регуляризатора, использующего двуязычные словари. В данной модели количество языков не ограничивается, однако тематическое распределение является общим для всех языков и представляет собой вероятностное распределение над словами всех языков. Кроме того, для приближения компонентов тематических распределений, соответствующих словам–переводам, используется квадратичный регуляризатор, который не вполне подходит для сравнения дискретных вероятностных распределений.

В настоящей статье предлагаются многоязычные модели с учетом двуязычных словарей, которые обобщаются на произвольное число языков, при этом для каждого языка строится своя матрица ϕ^ℓ_{wt},

$\ell \in L$. Для сравнения тематических распределений используется регуляризатор на основе дивергенции Кульбака–Лейблера, которая более адекватно оценивает сходство вероятностных распределений, чем квадратичная функция.

Предлагается несколько способов учета двуязычных словарей. Первый способ предполагает, что слова и их переводы относятся к одной теме с близкими вероятностями. Второй способ выражает вероятность слова в теме через вероятности его переводов в данной теме, а также через матрицу вероятностей появления такой пары в данной теме. Ниже будет дана математическая формализация этих методов. Обозначим через $\Pi_k(w) \subset W_k$ множество всех допустимых переводов слова $w \in W_\ell$, $\ell \in L$, на язык $k \in L$.

5.1. Приравнивание тематических профилей слов. Согласованность тематических распределений ϕ^ℓ_{wt}, $\ell \in L$, в разных языках можно оценить с помощью тематических профилей $p(t|w)$ и $p(t|u)$ для известных пар слов–переводов w и u. В модель вводится дополнительное требование: тематические профили $p(t|w)$ и $p(t|u)$ слова w и его перевода u должны быть покомпонентно близки.

Условие согласованности тематических распределений выражается через дивергенцию Кульбака–Лейблера между оценками $\widehat{p}(t|u,k)$ тематических профилей слов–переводов u и модельным тематическим профилем $p(t|w,\ell)$ слова w по всем словам всех языков:

$$\sum_{\ell\in L}\sum_{w\in W_\ell}\sum_{\substack{k\in L\\ k\neq\ell}}\sum_{u\in\Pi_k(w)} n^k_u\, KL\left(\widehat{p}(t|u,k)\,\middle\|\, p(t|w,\ell)\right) \to \min_{\Phi},$$

где $p(t|w,\ell) = p(w|t,\ell)\,\dfrac{p(t|\ell)}{p(w|\ell)} = \phi^\ell_{wt}\,\dfrac{p(t|\ell)}{p(w|\ell)}$, $\widehat{p}(t|u,k) = \dfrac{n^k_{ut}}{n^k_u}$, $n^k_u = \sum\limits_{t\in T} n^k_{ut}$.

Подставляя формулу для дивергенции Кульбака–Лейблера и исключая члены, не влияющие на результат максимизации, получаем следующий регуляризатор:

$$\mathcal{R}(\Phi,\Theta) = \sum_{\ell,k}\sum_{w\in W_\ell}\sum_{u\in\Pi_k(w)}\sum_{t\in T} n^k_{ut}\,\log\phi^\ell_{wt} \to \max_{\Phi,\Theta}.$$

Здесь суммирование $\sum\limits_{\ell,k}$ производится по всем языковым парам, имеющим словари.

В результате получаются формулы М-шага ЕМ-алгоритма (7), в которых к частотной оценке числителя ϕ^ℓ_{wt} в языке $\ell \in L$ добавляются частотные оценки по словам–переводам из других языков:

$$\phi^\ell_{wt} \propto n^\ell_{wt} + \tau \sum_{\substack{k\in L\\ k\neq\ell}}\sum_{u\in\Pi_k(w)} n^k_{ut}, \tag{10}$$

где τ — коэффициент регуляризации [11].

Рассмотренный метод обучения тематической модели с учетом словарей будем называть ML–TD (MultiLingual Translation Dictionary).

Такая модель имеет несколько недостатков. Во-первых, она притягивает тематические профили всех переводов одного и того же слова друг к другу, что может приводить к ошибкам в случае омонимии. Во-вторых, она не учитывает, что одно и то же слово может иметь различные переводы в разных темах.

5.2. Построение матрицы вероятностей переводов. Более реалистичным представляется предположение, что для каждого слова w языка ℓ вероятность его перевода u на язык k может зависеть от темы t. Это предположение легко формализуется с помощью регуляризатора, связывающего вероятность появления слова-перевода u в документе-переводе со словом w в документе-первоисточнике через матрицу вероятностей переводов $\pi^{kl}_{uwt} = p(u|w,t,k,\ell)$:

$$\phi^k_{ut} = p(u|t,k) = \sum_{w\in\Pi_\ell(u)} p(u|w,t,k,\ell)\,p(w|t,\ell) = \sum_{w\in\Pi_\ell(u)} \pi^{kl}_{uwt}\,\phi^\ell_{wt}. \tag{11}$$

Если пары (w,u) нет в словаре, то вероятности переводов равны нулю: $\pi^{k\ell}_{uwt} = \pi^{\ell k}_{wut} = 0$.

Потребуем, чтобы различия между оценками распределений $\widehat{p}(u|t,k) = \dfrac{n^k_{ut}}{n^k_t}$, полученными по коллекции, и модельными распределениями ϕ^k_{ut}, выраженными через матрицу переводов по формуле (11), были минимальными:

$$\sum_{\ell,k}\sum_{t\in T} n^k_t\, KL\left(\frac{n^k_{ut}}{n^k_t}\,\middle\|\,\sum_{w\in\Pi_\ell(u)} \pi^{k\ell}_{uwt}\,\phi^\ell_{wt}\right) \to \min_{\Phi,\Pi}.$$

Подставляя формулу для дивергенции Кульбака–Лейблера и исключая члены, не влияющие на решение задачи максимизации, получим выражение для регуляризатора:

$$\mathcal{R}(\Phi,\Theta,\Pi) = \sum_{\ell\in L}\sum_{\substack{k\in L\\ k\neq\ell}}\sum_{t\in T}\sum_{u\in W_k} n_{ut}^k \log \sum_{w\in\Pi_\ell(u)} \pi_{uwt}^{k\ell}\,\phi_{wt}^{\ell} \to \max_{\Phi,\Theta}.$$

Дискретные распределения $\pi_{uwt}^{k\ell}$ являются дополнительными параметрами тематической модели, которые тоже оцениваются по коллекции. На них накладываются ограничения неотрицательности и нормировки в дополнение к ограничениям (5):

$$\sum_{u\in W_k} \pi_{uwt}^{k\ell} = 1,\ \pi_{uwt}^{k\ell} \geqslant 0 \quad \forall\,\ell\in L,\ \forall\,k\in L,\ k\neq\ell,\ \forall\,w\in W_\ell,\ \forall\,t\in T.$$

При решении оптимизационной задачи для ϕ_{wt}^{ℓ} и $\pi_{uwt}^{k\ell}$ получаются следующие выражения М-шага ЕМ-алгоритма (7):

$$\phi_{wt}^{\ell} \propto n_{wt}^{\ell} + \tau \sum_{\substack{k\in L\\ k\neq\ell}} \sum_{u\in\Pi_k(w)} \pi_{wut}^{\ell k}\, n_{ut}^{k}; \quad \pi_{uwt}^{k\ell} \propto \pi_{wut}^{\ell k}\, n_{ut}^{k}. \tag{12}$$

Отметим, что для ϕ_{wt}^{ℓ} формулы (12) в методе, учитывающем вероятности переводов, и формулы (10) простого метода, основанного на приближении тематических профилей, отличаются лишь коэффициентом $\pi_{wut}^{\ell k}$. Данный способ учета словарей будем называть ML–TDP (MultiLingual Translation Dictionary Probability).

6. Вычислительные эксперименты.

6.1. Цели экспериментов. Основная цель проводимых экспериментов — продемонстрировать преимущество многоязычной модели, учитывающей параллельные или сравнимые документы совместно с двуязычными словарями переводов. Кроме того, мы показываем конкурентоспособность предлагаемого метода регуляризации по сравнению с уже существующими многоязычными моделями. Кроме того, в экспериментах сравниваются два описанных метода учета двуязычных словарей и исследуется зависимость модели от доли параллельных или сравнимых документов в обучении.

6.2. Текстовые коллекции. Построение многоязычных моделей проводилось на двух коллекциях. Параллельная коллекция Math состоит из математических статей на русском языке и их переводов на английский язык. Сравнимая коллекция Wiki содержит подмножество статей из категории “Математика” и связанных с ней категорий русской и английской Википедии, имеющих ссылки-интервики друг на друга.

Таблица 1

	Math, |D|	Math, |W|	Wiki, |D|	Wiki, |W|
Рус	154	4574	586	19305
En	154	6245	586	23413

На этапе предобработки из всех документов были удалены стоп-слова и произведена лемматизация. Характеристики русско-английских коллекций, использованных для многоязычного тематического моделирования, после предобработки приведены в табл. 1.

6.3. Словарь. Используемый в экспериментах двуязычный словарь был получен из русско-английского электронного словаря. Если слово имеет несколько переводов в целевом языке, то рассматривались все переводы, причем в словарь отбирались только однословные переводы, переводы-словосочетания не учитывались. Получившийся словарь включает 82 642 пары переводов.

6.4. Обучение модели. Во всех экспериментах использовался регуляризованный ЕМ-алгоритм, модифицированный для многоязычных моделей. Во всех экспериментах 100 итераций было достаточно для сходимости моделей.

При обучении модели по обучающей выборке настраиваются тематические матрицы ϕ_{wt}^{ℓ}, $\ell \in L$. Для всех документов как обучения, так и контроля тематические профили θ_{td} получаются с помощью Е-шага ЕМ-алгоритма, тематические матрицы ϕ_{wt}^{ℓ}, $\ell \in L$, при этом не изменяются.

6.5. Выбор числа тем. Эксперименты проводились для числа тем, равного 25, 50 и 100. При числе тем, меньшем 25, качество построенной тематической модели не удовлетворительно. Рассмотрение числа тем, большего 100, представляется нецелесообразным из-за небольшого объема используемых коллекций. Выбранные значения позволяют исследовать влияние числа тем на качество модели.

6.6. Критерии качества. Одним из интересующих нас приложений тематического моделирования является тематический поиск для рекомендации документов, близких по тематике к данному. Частным

случаем является кросс-язычный поиск, когда запросом является документ на одном языке, а поиск производится среди документов другого языка. Требуется найти документы на другом языке, максимально похожие на запрос по тематическому профилю. В нашей работе, как и в работах [6, 7], кросс-язычный поиск является не самоцелью, а способом сравнения многоязычных тематических моделей, который дает более интерпретируемые результаты, чем распространенная в тематическом моделировании мера качества — перплексия. При кросс-язычном поиске для каждого документа-запроса $q_\ell \in Q$ на исходном языке ℓ документы $d_k \in S$ на целевом языке k ранжируются по возрастанию расстояния Хеллингера между их тематическими профилями: $H(q_\ell, d_k) = \frac{1}{\sqrt{2}}\sqrt{\sum_{t\in T}\left(\sqrt{\theta^\ell_{tq}} - \sqrt{\theta^k_{td}}\right)^2}$.

Так как для каждого запроса q найдется его перевод в множестве S, то мерой качества может служить позиция $\mathrm{Pos}\,(d, q, S)$ истинного документа-перевода d в ранжированном списке. Этот показатель усреднялся по всем запросам из множества Q: $\mathrm{AveragePos}\,(Q) = \frac{1}{|Q|}\sum_{q\in Q}\mathrm{Pos}\,(d, q, S)$.

Для получения более точных оценок применялся скользящий контроль с 10 блоками. Для каждого разбиения коллекции на обучение и контроль запросами являлись документы обучения или контроля, в качестве множества доступных документов для поиска перевода были взяты все имеющиеся документы на другом языке.

6.7. Сравнение с другими реализациями многоязычных тематических моделей. В программном пакете MALLET (MAchine Learning for LanguagE Toolkit) [24] реализована многоязычная тематическая модель на базе модели PLTM [7], использующая параллельные коллекции для связывания языков. Для обучения многоязычной модели были взяты параметры по умолчанию. С помощью обученной модели были получены распределения тем в документах на обучении и контроле, после чего оценивалось качество кросс-язычного поиска.

Таблица 2

Средняя позиция кросс-язычного поиска для моделей ML–P, ARTM и PLTM, MALLET

Модель		25 тем		50 тем		100 тем	
		Обучение	Контроль	Обучение	Контроль	Обучение	Контроль
Math	ML–P, ARTM	1.031	**1.051**	1.007	**1.002**	1.007	1.009
	PLTM, MALLET	**1.021**	1.066	**1.002**	1.041	**1.000**	**1.006**
Wiki	ML–P, ARTM	2.578	**4.117**	1.806	**2.826**	1.423	**2.186**
	PLTM, MALLET	**2.261**	4.596	**1.586**	3.489	**1.236**	2.693

Эксперимент проводился на коллекциях Math и Wiki. Сравнивались модели PLTM (MALLET) и ML–P. Результаты сравнения приведены в табл. 2. В каждой ячейке жирным шрифтом выделен лучший результат. Как следует из этой таблицы, на коллекции Math малого объема модели почти не различаются. На более объемной коллекции Wiki модель PLTM, реализованная в программном пакете MALLET, лучше, чем модель ML–P, на обучении, но хуже на контроле. В целом, можно сказать, что данные модели показывают сравнимое качество кросс-язычного поиска.

6.8. Сравнение методов обучения многоязычных моделей с использованием двуязычного словаря. Эксперимент проводился на коллекциях Math и Wiki. Информация о связи документов-переводов не учитывалась при обучении модели. Сравнивались два метода учета словарей: ML–TD и ML–TDP.

Метод ML–TDP показывает лучшее качество кросс-язычного поиска (табл. 3), т.е. дополнительный учет тематики переводов способен улучшать модель.

Кроме того, было проведено сравнение с методом поиска документов-переводов, основанным на векторном представлении документов в пространстве слов из словаря языка. Для кросс-язычного поиска запрос пословно переводится на целевой язык при помощи двуязычного словаря, после чего рассчитывается близость вектора-запроса к векторам документов в пространстве слов целевого языка.

Сравнивались векторные представления документов на основе статистик tf (term frequency) и tf-idf (term frequency-inverse document frequency), в качестве мер близости использовались косинусная метрика и расстояние Хеллингера. Для улучшения качества поиска в документах и запросах были оставлены топ-N слов, соответствующих самым большим значениям статистики, все остальные обнулялись. После этого

сокращения запросы переводились на целевой язык, затем производился поиск максимально похожего документа среди всех документов целевого языка.

Таблица 3

Средняя позиция кросс-язычного поиска при обучении с помощью словарей

Модель		25 тем		50 тем		100 тем	
		Обучение	Контроль	Обучение	Контроль	Обучение	Контроль
Math	Vector	**1.000**					
	ML–TD	8.690	5.762	5.162	4.848	2.836	1.318
	ML–TDP	2.990	**1.646**	1.955	**1.091**	1.470	**1.006**
Wiki	Vector	**4.898**					
	ML–TD	47.898	41.018	38.827	13.29	29.167	7.269
	ML–TDP	40.215	**32.160**	33.230	**9.945**	24.100	**5.876**

Сочетание статистики tf-idf, топ-200 слов и расстояния Хеллингера позволяет достичь наилучших результатов. В табл. 3 данный метод обозначен как Vector. По качеству кросс-язычного поиска он превосходит тематический поиск с моделями, обученными только по двуязычным словарям переводов.

Таблица 4

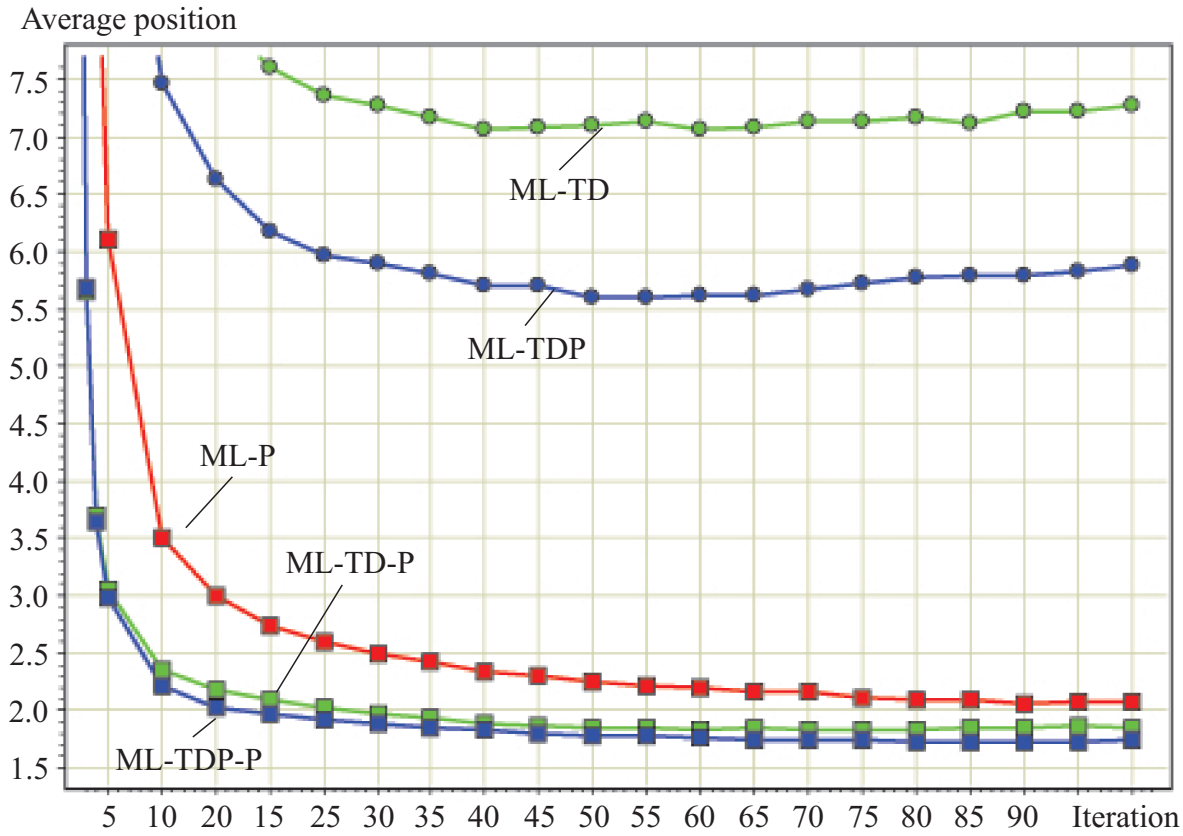
Средняя позиция кросс-язычного поиска на коллекции Wiki

Модель	25 тем		50 тем		100 тем	
	Обучение	Контроль	Обучение	Контроль	Обучение	Контроль
ML–P	2.578	**4.117**	1.806	**2.826**	1.423	**2.186**
ML–TDP	40.215	32.160	33.230	9.945	24.100	5.876
ML–TDP–P0.1	32.494	12.315	26.073	8.045	20.153	4.850
ML–TDP–P0.2	24.852	9.737	20.976	6.254	15.732	3.794
ML–TDP–P0.3	19.777	7.487	16.226	4.895	13.717	3.567
ML–TDP–P0.4	14.859	6.098	12.873	4.379	11.000	3.030
ML–TDP–P0.5	12.292	5.461	9.412	3.508	8.528	2.675
ML–TDP–P0.6	9.179	4.767	7.720	3.349	6.618	2.637
ML–TDP–P0.7	7.464	**4.052**	6.213	2.977	5.081	**2.150**
ML–TDP–P0.8	5.620	**4.027**	4.377	**2.674**	3.673	**2.058**
ML–TDP–P0.9	4.102	**3.701**	3.156	**2.555**	2.545	**1.886**
ML–TDP–P	2.528	**3.578**	1.713	**2.265**	1.373	**1.757**

6.9. Зависимость качества поиска от числа связанных документов в обучающей выборке. Эксперимент проводился на коллекции Wiki. Коллекция Math не рассматривалась в силу малого размера. Доля документов, для которых учитывались связи, изменялась от 0.0 до 1.0 с шагом 0.1. Для каждого значения параметра связанности проводилось два эксперимента: с использованием информации о связи слов-переводов в двуязычном словаре и без этой информации. В качестве метода учета словарей был выбран метод ML–TDP с настройкой вероятностей переводов как показавший лучшие результаты, чем более простой метод ML–TD. Для именования моделей, в которых учитываются связи между документами, к основному префиксу ML–P добавляется указание о доле связанных документов при обучении, например, ML–P0.3 (MultiLingual Parallel 0.3). Если дополнительно использовался двуязычный словарь, это также указывается: ML–TDP–P0.3.

Из табл. 4 видно, что с увеличением доли параллельных документов на обучении качество кросс-язычного поиска монотонно увеличивается как на обучающей выборке, так и на контрольной. Кроме того, из двух первых строк таблицы видно, что сравнимая коллекция является более сильным источником многоязычной информации, чем двуязычный словарь переводов. Последние строки таблицы показывают, что одновременный учет даже не всех связанных документов и двуязычных словарей может дать улучшение по сравнению с методами, использующими только один источник информации. Комбинация параллельной или сравнимой коллекции и двуязычных словарей дает наилучшее качество кросс-язычного поиска.

На графике (см. рисунок) представлено качество поиска для коллекции Wiki со 100 темами по итерациям EM-алгоритма для моделей, построенных с использованием всей параллельной коллекции, двуязычного словаря или обоих методов. Сходимость наблюдается начиная примерно с 50-й итерации. Сочетание параллельности коллекции и двуязычного словаря улучшает качество поиска по сравнению с каким-то одним методом.



Качество кросс-язычного поиска для коллекции **Wiki** со 100 темами

6.10. Анализ найденных переводов в различных темах. Метод ML–TDP строит вероятностные распределения $p(u|w,t)$ появления слова u из языка k в качестве перевода слова w из языка ℓ в теме t. В качестве начального значения $p(u|w,t)$ для всех тем берется величина, обратная числу переводов слова w: $p(u|w,t) = \dfrac{[u \in \Pi_k(w)]}{|\Pi_k(w)|}$, что соответствует равномерному распределению вероятности на множестве слов-переводов $\Pi_k(w)$. В процессе обучения модели и настройки тематических распределений $p(w|t,\ell)$ матрица вероятностей переводов для различных тем и пар слов действительно изменяется по сравнению с начальной инициализацией. Следующие примеры показывают, что для разных тем определяются свои, более специфичные переводы слов.

Слово “сумма” имеет переводы “sum” и “total”. Перевод “sum” имеет вероятность больше 0.9 в темах 6, 12, 20. Перевод “total” наиболее вероятен в темах 5, 19, 22. Список верхних 10 слов на русском и английском языках для данных тем представлен в табл. 5.

Темы 6, 12 и 20 имеют математическую направленность. Пара “сумма”–“sum” используется в них в значении арифметической суммы. Темы 5, 19, 22 различаются по содержанию, судя по топ-10 слов тему 5 можно отнести к космической тематике, тему 19 — к компьютерной, тему 22 — к игровой. Пара “сумма”–“total” употребляется в данных темах в значении общего или итогового количества.

7. Заключение. В настоящей статье предложен метод обучения многоязычных тематических моделей на основе аддитивной регуляризации (ARTM). В рамках многоязычных тематических моделей регуляризаторы содержат дополнительную информацию, позволяющую связывать слова и документы различных языков для построения согласованных тематических распределений. Рассмотрено два основных источника многоязычной информации: параллельные и сравнимые коллекции и двуязычные словари переводов.

Таблица 5

Топ-10 слов в темах, где вероятность перевода “сумма”–“sum” более 0.9 (темы 6, 12, 20) либо вероятность перевода “сумма”–“total” более 0.9 (темы 5, 19, 22)

Тема 6		Тема 12		Тема 20	
множество	set	математика	triangle	вектор	vector
пространство	space	треугольник	square	координата	coordinate
группа	point	теорема	number	пространство	field
точка	left	точка	point	преобразование	tensor
элемент	limit	математический	theorem	базис	transform
функция	symmetry	угол	angle	тензор	basis
предел	function	координата	mathematics	сила	space
отображение	open	экономика	real	векторный	force
симметрия	property	число	theory	точка	rotation
открытый	topology	квадрат	geometry	система	thermometer
Тема 5		**Тема 19**		**Тема 22**	
орбита	space	программный	software	игра	game
аппарат	nasum	версия	version	видеосигнал	character
космический	orbit	работа	news	игрок	video
земля	instrument	компания	company	фильм	player
поверхность	earth	анонимный	work	головоломка	series
солнечный	surface	примечание	note	серия	puzzle
станция	solar	терминатор	release	качество	movie
запуск	system	журнал	support	шахматы	jason
система	landing	рей	terminator	джейсон	world
атмосфера	camera	персонаж	anonymous	буква	chess

Сравнение различных методов учета многоязычной информации в тематических моделях показало, что комбинирование параллельных или сравнимых документов и двуязычных словарей улучшает качество кросс-язычного поиска в сравнении с моделями, использующими только один источник информации. В дальнейших исследованиях планируется проверить данные результаты на более объемных коллекциях.

При использовании двуязычных словарей оценивание вероятностей переводов в зависимости от темы улучшает качество кросс-язычного поиска по сравнению с более простым методом. Кроме того, в этом случае для каждого перевода в словаре тематическая модель позволяет указать контекст (набор тем), в котором данный перевод наиболее уместен.

Пословный перевод запроса на целевой язык с помощью двуязычного словаря и последующий поиск в векторном пространстве слов целевого языка дает лучшие результаты, чем тематический поиск, что указывает на потенциальную возможность дальнейшего улучшения тематических моделей, использующих двуязычные словари переводов.

Работа выполнена при финансовой поддержке РФФИ (коды проектов 14–07–31176 и 14–07–00908).

СПИСОК ЛИТЕРАТУРЫ

1. *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’99. New York: ACM Press, 1999. 50–57.
2. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet allocation // J. Mach. Learn. Res. 2003. **3**. 993–1022.
3. *Boyd-Graber J., Blei D.M.* Multilingual topic models for unaligned text // Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI’09. Arlington: AUAI Press, 2009. 75–82.
4. *Jagarlamudi J., Daumé H. III, Udupa R.* From bilingual dictionaries to interlingual document representations // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT’11. Stroudsburg: ACL Press, 2011. 147–152.
5. *Ni X., Sun J.-T., Hu J., Chen Z.* Cross lingual text classification by mining multilingual topics from Wikipedia // Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM’11. New York: ACM Press, 375–384.
6. *Ni X., Sun J.-T., Hu J., Chen Z.* Mining multilingual topics from Wikipedia // Proceedings of the 18th International Conference on World Wide Web, WWW’09. New York: ACM Press, 2009. 1155–1156.
7. *Mimno D., Wallach H.M., Naradowsky J., Smith D.A., McCallum A.* Polylingual topic models // Proceedings of

the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP'09. Stroudsburg: ACL Press, 2009. 880–889.
8. *Smet W.D., Moens M.-F.* Cross-language linking of news stories on the web using interlingual topic modelling // Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, SWSM'09. New York: ACM Press, 2009. 57–64.
9. *Zhang D., Mei Q., Zhai C.X.* Cross-lingual latent topic extraction // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL'10. Stroudsburg: ACL Press, 2010. 1128–1137.
10. *Boyd-Graber J., Resnik P.* Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP'10. Stroudsburg: ACL Press, 2010. 45–55.
11. *Воронцов К.В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады Академии наук. 2014. **456**, № 3. 268–271.
12. *Vorontsov K.V., Potapenko A.A.* Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization // Analysis of Images, Social networks and Texts. Communications in Computer and Information Science. Vol. 436. Heidelberg: Springer, 2014. 29–46.
13. *Воронцов К.В., Потапенко А.А.* Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии. Вып. 13. М.: Изд-во РГГУ, 2014. 676–687.
14. *Si L., Jin R.* Adjusting mixture weights of Gaussian mixture model via regularized probabilistic latent semantic analysis // Lecture Notes in Computer Science. Vol. 3518. Heidelberg: Springer, 2005. 622–631.
15. *Chien J.-T., Wu M.-S.* Adaptive Bayesian latent semantic analysis // Trans. Audio, Speech and Lang. Proc. 2008. **16**, N 1. 198–207.
16. *Mei Q., Cai D., Zhang D., Zhai C.X.* Topic modeling with network regularization // In Proceedings of the 17th International Conference on World Wide Web, WWW'08. New York: ACM Press, 2008. 101–110.
17. *Wang Q., Xu J., Li H., Craswell N.* Regularized latent semantic indexing // Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'11. New York: ACM Press, 2011. 685–694.
18. *Koehn P.* Europarl: a parallel corpus for statistical machine translation // Conference Proc. 10th Machine Translation Summit. 2005 (http://www.mt-archive.info/MTS-2005-Koehn.pdf/).
19. *Miller G.A.* WordNet: a lexical database for English // Commun. ACM. 1995. **38**, N 11. 39–41.
20. *Vossen P.* EuroWordNet: a multilingual database with lexical semantic networks. Norwell: Kluwer Academic Publishers, 1998.
21. *Navigli R., Ponzetto S.P.* BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network // Artif. Intell. 2012. **193**. 217–250.
22. *Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P.N., Hellmann S., Morsey M., van Kleef P., Auer S., Bizer C.* DBpedia — a large-scale, multilingual knowledge base extracted from Wikipedia // Semantic Web Journal. 2014 (http://www.semantic-web-journal.net/system/files/swj499.pdf).
23. *de Melo G., Weikum G.* MENTA: inducing multilingual taxonomies from Wikipedia // Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10. New York: ACM Press, 2010. 1099–1108.
24. *McCallum A.K.* MALLET: a machine learning for language toolkit. 2002 (http://mallet.cs.umass.edu).

Поступила в редакцию
27.11.2014

Regularization of Multilingual Topic Models

M. A. Dudarenko [1]

[1] *Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics; Leninskie Gory, Moscow, 119992, Russia; Graduate Student, e-mail: m.dudarenko@gmail.com*

Received November 27, 2014

Abstract: A multilingual probabilistic topic model based on the additive regularization ARTM allowing to combine both a parallel or comparable corpus and a bilingual translation dictionary is proposed. Two approaches to include information from a bilingual dictionary are discussed: the first one takes into account only the fact of connection between word translations, whereas the second one learns the translation probabilities for each topic. To measure the quality of the proposed multilingual topic model, a cross-language search is performed.

For each query document in one language, it is found its translation on an other language. It is shown that the combined translation of words from a bilingual dictionary and the corresponding connected documents improves the cross-lingual search compared to the models using only one information source. The use of learning word translation probabilities for bilingual dictionaries improves the quality of the model and allows one to determine a context (a set of topics) for each pair of word translations, where these translations are appropriate.

Keywords: multilingual topic model, probabilistic topic model, parallel corpus, comparable corpus, bilingual dictionary, regularization, cross-language search.

References

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

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

3. J. Boyd-Graber and D. M. Blei, "Multilingual Topic Models for Unaligned Text," in *Proc. 25th Conf. on Uncertainty in Artificial Intelligence, Montreal, June 18–21, 2009* (AUAI Press, Arlington, 2009), pp. 75–82.

4. J. Jagarlamudi, H. Daumé, and R. Udupa, "From Bilingual Dictionaries to Interlingual Document Representations," in *Proc. 49th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, Portland, June 19–24, 2011* (ACL Press, Stroudsburg, 2011), Vol. 2, pp. 147–152.

5. X. Ni, J.-T. Sun, J. Hu, and Z. Chen, "Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia," in *Proc. 4th ACM Int. Conf. on Web Search and Data Mining, Hong Kong, February 9–12, 2011* (ACM Press, New York, 2011), pp. 375–384.

6. X. Ni, J.-T. Sun, J. Hu, and Z. Chen, "Mining Multilingual Topics from Wikipedia," in *Proc. 18th ACM Int. Conf. on World Wide Web, Madrid, April 20–24, 2009* (ACM Press, New York, 2009), pp. 1155–1156.

7. D. Mimno, H. M. Wallach, J. Naradowsky, et al., "Polylingual Topic Models," in *Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, Singapore, August 6–7, 2009* (ACL Press, Stroudsburg, 2009), Vol. 2, pp. 880–889.

8. W. De Smet and M.-F. Moens, "Cross-Language Linking of News Stories on the Web Using Interlingual Topic Modelling," in *Proc. 2nd ACM Workshop on Social Web Search and Mining, Hong Kong, November 2, 2009* (ACM Press, New York, 2009), pp. 57–64.

9. D. Zhang, Q. Mei, and C. X. Zhai, "Cross-Lingual Latent Topic Extraction," in *Proc. 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, July 11-16, 2010* (ACL Press, Stroudsburg, 2010), pp. 1128–1137.

10. J. Boyd-Graber and P. Resnik, "Holistic Sentiment Analysis across Languages: Multilingual Supervised Latent Dirichlet Allocation," in *Proc. 2010 Conf. on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts, October 9–11, 2010* (ACL Press, Stroudsburg, 2010), pp. 45–55.

11. K. V. Vorontsov, "Additive Regularization for Topic Models of Text Collections," Dokl. Akad. Nauk **456** (3), 268–271 (2014) [Dokl. Math. **89** (3), 301–304 (2014)].

12. K. Vorontsov and A. Potapenko, "Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization," in *Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science* (Springer, Heidelberg, 2014), Vol. 436, pp. 29–46.

13. K. V. Vorontsov and A. A. Potapenko, "Regularization of Probabilistic Topic Models to Improve Interpretability and Determine the Number of Topics," in *Computational Linguistics and Intellectual Technologies* (Ross. Gos. Gumanitarn. Univ., Moscow, 2014), Issue 13, pp. 676–687.

14. L. Si and R. Jin, "Adjusting Mixture Weights of Gaussian Mixture Model via Regularized Probabilistic Latent Semantic Analysis," in *Lecture Notes in Computer Science* (Springer, Heidelberg, 2005), Vol. 3518, pp. 622–631.

15. J.-T. Chien and M.-S. Wu, "Adaptive Bayesian Latent Semantic Analysis," IEEE Trans. Audio, Speech and Lang. Proc. **16** (1), 198–207 (2008).

16. Q. Mei, D. Cai, D. Zhang, C.X. Zhai, "Topic Modeling with Network Regularization," in *Proc. 17th Int. Conf. on World Wide Web, Beijing, April 21-25, 2008* (ACM Press, New York, 2008), pp. 101–110.

17. Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized Latent Semantic Indexing," in *Proc. 34th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Beijing, July 24–28* (ACM Press, New York, 2011), pp. 685–694.

18. P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Proc. 10th Machine*

Translation Summit, Phuket, Thailand, September 12–16, 2005.
http://www.mt-archive.info/MTS-2005-Koehn.pdf/. Cited January 7, 2015.

19. G. A. Miller, "WordNet: A Lexical Database for English," Commun. ACM **38** (11), 39–41 (1995).

20. P. Vossen, *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* (Kluwer, Norwell, 1998).

21. R. Navigli and S. P. Ponzetto, "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network," Artif. Intell. **193**, 217–250 (2012).

22. J. Lehmann, R. Isele, M. Jakob, et al., "DBpedia — A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia," Semantic Web Journal (2014).
http://www.semantic-web-journal.net/system/files/swj499.pdf. Cited January 7, 2015.

23. G. de Melo and G. Weikum, "MENTA: Inducing Multilingual Taxonomies from Wikipedia," in *Proc. 19th ACM International Conference on Information and Knowledge Management, Toronto, October 26–30* (ACM Press, New York, 2010), pp. 1099–1108.

24. A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu. Cited January 7, 2015.