Регуляризация многоязычной тематической модели на основе параллельной коллекции и двуязычного словаря.

Дударенко Марина Алексеевна

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: m.dudarenko@gmail.com

Вероятностное тематическое моделирование используется для извлечения латентных тем t из текстовых коллекций. Каждый документ d представляется вероятностным распределением $p(t|d)$ на множестве тем t. Каждая тема t, в свою очередь, представляется вероятностным распределением $p(w|t)$ на множестве слов w из словаря коллекции. Наиболее известны тематические модели PLSA [2] и LDA [3].

В многоязычных коллекциях каждый документ написан на своем языке, поэтому одноязычная модель не может объединять слова из разных языков в одну тему, опираясь на статистику их совместного появления в документах. Для получения многоязычных согласованных тем нужны дополнительные ресурсы, связывающие слова из разных языков. Это могут быть параллельные и сравнимые коллекции или двуязычные словари и структурированные базы знаний (WordNet, MENTA), в которых проставлены кросс-язычные связи между документами или словами соответственно.

Большинство многоязычных моделей основываются на модели LDA и учитывают либо связанные документы [4], либо двуязычные словари [5]. В данной работе предлагается многоязычная модель на основе аддитивной регуляризации ARTM [1], которая позволяет одновременно учесть как связи между документами параллельной или сравнимой коллекции, так и двуязычный словарь. Предлагается два способа использования информации из двуязычного словаря. В первом способе регуляризатор сближает распределения тем $p(t|w)$ и $p(t|u)$ для слов переводов w и u. Во втором способе строится матрица вероятностей $p(u|w,t)$ появления перевода u слова w в теме t, через которую регуляризатор сближает распределения $p(w|t)$ и $p(u|t)$.

Оценка моделей производилась на двух коллекциях. Параллельная коллекция **Math** содержит 154 математических статьи на русском языке и их переводы на английский язык. Сравнимая коллекция **Wiki** получена из 586 статей русской Википедии, включенных в категорию «Математика» и связанных кросс-язычными ссылками

Таблица 1: Средняя позиция документа–перевода в кросс-язычном поиске для моделей, учитывающих различные источники многоязычной информации.

Источник многоязычной информации	**Math**	**Wiki**
Словарь (1)	5.6	26.3
Словарь (2)	2.0	22.2
Параллельные документы (0.8)	2.4	10.4
Параллельные документы (1.0)	1.6	6.2
Словарь (1) и параллельные документы (0.8)	1.7	7.1
Словарь (2) и параллельные документы (0.8)	1.2	5.5
Словарь (1) и параллельные документы (1.0)	1.2	5.2
Словарь (2) и параллельные документы (1.0)	1.1	4.7

с их английскими эквивалентами. После удаления стоп-слов и лемматизации словари коллекции **Math** насчитывают 4574 и 6245 слов, коллекции **Wiki** — 19305 и 28804 слов для русского и английского языков соответственно.

Качество многоязычных моделей оценивалось по средней позиции документа–перевода в поисковой выдаче кросс-язычного поиска. В этом случае запросом является документ на одном языке, а поиск производится среди документов другого языка. В качестве запросов использовались как целые документы, так и их фрагменты.

Было проведено сравнение двух способов включения двуязычного словаря в модель. Также сравнивались модели, использующие параллельную коллекцию, двуязычный словарь и их сочетание. Доля документов параллельной коллекции, у которых проставлены связи во время обучения, варьировалась от 0 до 1. Для всех экспериментов число тем равно 50.

Показано, что при использовании двуязычного словаря учет матрицы вероятностей переводов слов $p(u|w,t)$ дает лучшие результаты. Добавление информации из двуязычного словаря улучшает качество кросс-язычного поиска при любом количестве связанных документов. Кроме того, модель, у которой были проставлены все связи между документами, но не использовался словарь, оказалась хуже некоторых моделей, использующих лишь долю связей и двуязычный словарь с построением матрицы $p(u|w,t)$. Значения средней позиции документа–перевода в выдаче приведены в таблице . В скобках для словаря указан способ учета слов–переводов, для параллель-

ных документов — доля проставленных связей между документами–переводами.

Таким образом, комбинированный учет информации о словах-переводах из двуязычного словаря в моделях, основывающихся на параллельных или связанных коллекциях, улучшает качество кросс-язычного поиска. При использовании двуязычных словарей включение матрицы вероятностей появления пары слов-переводов в конкретной теме также позволяет улучшить модель.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты 14-07-31176, 14-07-00908.

Литература

1. Воронцов К. В. Аддитивная регуляризация тематических моделей текстовых коллекций // Доклады Академии наук, Т. 499, № 3, 2014.

2. Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA, 1999, P. 50–57.

3. Blei D. M., Ng A. Y., and Jordan M. I. Latent dirichlet allocation // In J. Mach. Learn. Res. 3. 2003, P. 993–1022.

4. Ni X., Sun J. T., Hu J., and Chen Z. Mining multilingual topics from Wikipedia // Proceedings of the 18th international conference on World wide web. New York, USA, 2009, P. 1155–1156.

5. Jagarlamudi J. and Daumé H., III. From bilingual dictionaries to interlingual document representations // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Vol. 2. Stroudsburg, PA, USA, 2011, P. 147–152.