Тематические модели для построения интерпретируемых векторных представлений слов

Попов Артём Сергеевич

Студент

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

E-mail: artems-07@mail.ru

Тематическое моделирование (topic modeling) и векторные представления слов (word embedding) — это два различных подхода к автоматическому выявлению смыслов в текстовых коллекциях. Оба подхода применяются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

Большинство моделей для построения векторных представлений слов основано на использовании гипотезы дистрибутивности, заключающейся в том, что слова, встречающиеся в схожих контекстах, имеют семантически близкие значения. Самой известной моделью такого типа является модель SGNS, или word2vec [1], позволяющая эффективно решать задачи поиска семантически близких слов и выявления аналогий в парах слов. Полученные этой моделью вектора являются плотными, а компоненты векторов неинтерпретируемыми.

В основе вероятностных тематических моделей обычно лежит другая гипотеза — независимости, заключающаяся в том, что порядок слов в документах коллекции не важен для выявления его тематики. В вероятностных тематических моделях слова w и документы d описываются дискретными распределениями $p(t|w)$, $p(t|d)$ на множестве тем $t \in T$, то есть неотрицательными нормированными векторами размерности $|T|$.

Компоненты в тематических векторных представлениях слов являются интерпретируемыми. Кроме того, они разрежены, поскольку каждое слово относится, как правило, к небольшому числу тем. В то же время, они сильно проигрывают по качеству решения задач близости и аналогии моделям, основанным на гипотезе дистрибутивности.

Целью данной работы является объединение достоинств обоих подходов и построение модели, в которой векторные представления слов позволяли бы решать задачи аналогии и близости, но при этом их компоненты оставались бы вероятностными, разреженными и интерпретируемыми. Для этого предлагается использовать тематическую модель ARTM, которая описывает многие тематические моде-

ли в рамках единого формализма и позволяет их комбинировать. Для формализации гипотезы дистрибутивности по каждому слову строится псевдо-документ, представляющий собой объединение всех локальных контекстов данного слова в коллекции [3].

Вычислительные эксперименты проводились на коллекции статей Википедии 2016 года. Предварительно отбрасывалось 25 стоп-слов, а затем словарь сокращался до 100 000 самых частотных слов оставшейся части коллекции. При перенарезке коллекции использовался сабсэмплинг, слова, встречавшиеся друг с другом менее 5 раз не учитывались при построении модели. На полученных совстречаемостях было обучено несколько моделей ARTM и SGNS с различными гиперпараметрами. Также, на исходной коллекции была обучена стандартная тематическая модель LDA.

Модель ARTM, построенная по псевдо-документам, по качеству решения задачи близости заметно превосходит LDA и сопоставима с моделью SGNS. Сравнение различных способов измерения близости между тематическими векторами показало, что косинусное расстояние и расстояние Хеллингера уступают скалярному произведению векторов. В то же время вектора, полученные с помощью тематических моделей по псевдо-документам, остаются разреженными, а их компоненты интерпретируемыми. Разреженность и интерпретируемость можно контролировать с помощью механизма регуляризации ARTM, не ухудшая качество решения задачи близости слов.

Литература

1. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 2013, P. 3111–3119.
2. Vorontsov K., Potapenko A. Additive regularization of topic models // Machine Learning, 2015, V. 101, №1, P. 303–323.
3. Zuo Y., Zhao J., Xu K. Word Network Topic Model: a simple but general solution for short and imbalanced texts // Knowledge and Information Systems, 2016, V. 48, №2, P. 379–398.