Рецензия на доклад Анны Потапенко «Вероятностное тематическое моделирование и нейросетевые модели языка для обучения векторного представления слов, контекстов и документов»

Доклад посвящен исследованиям в области объединения двух подходов к обучению векторных моделей языка: вероятностных и нейросетевых. Рассказ состоял как из обзора основополагающих работ в данной области, так и из собственных результатов автора, что говорит о вовлеченности в процесс и не может не радовать. Выступление было хорошо подготовлено, актуальность задачи также не вызывает сомнений, о чем можно судить по достижениям автора в прошлом году и планам на будущее.

В первой части были рассмотрены два варианта обучения векторных моделей языка: вероятностные (LDA, PLSA) и нейросетевые (N-gram language models, word2vec, GloVe), а также области их применения. Были приведены отличия рассматриваемых подходов: различные способы моделирования вероятности, различные методы оптимизации... С другой стороны, было показано, что при различном определении «документа» для тематических моделей, они могут учить матричные разложения, эквивалентные обучаемым нейросетевыми моделями. Это позволяет объединить преимущества вероятностных и нейросетевых подходов: интерпретируемость и сохранение семантической близости в векторном пространстве соответственно.

Во второй части были приведены первые полученные результаты гибридной модели. Рассматривались задачи семантической близости и аналогий. На датасете WordSim353 было показано качество, сравнимое с word2vec и GloVe, что очень здорово для первых экспериментов с новой моделью, несмотря на сильно ограниченный размер тестовой выборки. К сожалению, из доклада мне не удалось четко понять, насколько производителен EM-алгоритм для оптимизации тематических моделей по сравнению с online-SGD у word2vec. Возможно, этот момент стоило осветить немного подробнее. Затем были приведены наиболее вероятные темы для некоторых векторов из обученного пространства. Выглядит интересно, однако на слайдах было бы здорово помимо списка тематик увидеть и соответствующий им список вероятностей, чтобы можно было оценить интерпретируемость полученной модели. Также была рассмотрена реальная задача построения динамической тематической модели британского искусства. Предлагаемый алгоритм опять заметно превосходит базовую модель PLSA, причем он еще может быть улучшен за счет совместного обучения на данных слова-слова и слова-документы.

В завершающей части был приведен ряд наиболее значимых достижений докладчика за последний год, а также планов на будущее. Настолько внушительный список не позволяет сомневаться в компетентности автора и предстоящих серьезных достижениях в рассматриваемой области.

21 февраля 2016 г. Олег Найдин