Рецензия на доклад Потапенко Анны на тему:
«Возможности использования локального контекста в вероятностных тематических моделях»

Рецензент: Лобачева Екатерина

Данный доклад представляет собой обзор различных возможностей учета последовательности слов в документах при построении тематических моделей. Конкретнее, рассматриваются методы учета локального контекста слов. В целом доклад был очень интересен, так как в нем проводились параллели между различными современными методами, использующимися в лингвистическом моделировании. Однако на данный момент времени докладчица явно находится в активном поиске направлений развития своего исследования, поэтому доклад недостаточно систематизирован, не до конца понятны связи между всеми упомянутыми моделями. Думаю, в дальнейшем тема конкретизируется и эта проблема исчезнет сама собой. Далее я кратко опишу основные идеи доклада и прокомментирую их в меру своего понимания.

Тематическая модель на «псевдодокументах»
Предлагается создать набор «псевдодокументов»: для каждого слова w создается документ из всех слов, которые встречаются в контексте длины k с данным. На таких документах можно обучать как стандартные тематические модели, например, PLSA, так и контекстные модели Word2Vec и различные его модификации.

Мне понравилась аналогия между тематическими моделями и моделями контекста: и то, и другое, по сути, представляет собой разложение матрицы в произведение двух. При этом, однако, для этих моделей по-разному задаются вероятности и оптимизируются разные функционалы. Интересно было бы посмотреть на результаты сравнения этих подходов, а также понять, как влияет оптимизируемый функционал на итоговое решение.
Также, возможно, стоит подумать не только о моделях локального контекста, но и об учете более продолжительных зависимостей в данных с помощью, например, скрытых марковских моделей или рекуррентных нейронных сетей. Однако это может быть уже ненужным усложнением для задачи тематического моделирования.

Мультимодальная тематическая модель
Предлагается реализовать мультимодальную тематическую модель, в которую войдет обычная тематическая модель на исходных документах и какая-то из моделей из предыдущего пункта, позволяющая учитывать локальный контекст.

Идея кажется простой, уместной и, думаю, перспективной.

Работа с короткими документами
Применяется идея о применении контекстных моделей на основе матрицы PMI в тематическом моделировании на коротких текстах.

Я не поняла, что в данном случае берется за документ: одна запись или какой-то более большой блок? Пробовалось ли обучать модель на больших текстах с особой регуляризацией (слова, которые часто встречаются в коротких текстах, должны быть основой тем, обученных по длинным текстам) с последующим применением к коротким текстам?

PMI-регуляризация тематической модели
Предлагается метод аддитивной регуляризации для учета локального контекста с помощью PMI-матриц. Этот метод позволяет сделать смену тем внутри документа более плавной, то есть приводит к тому, что документ представляет собой не мешанину слов из разных тем, а последовательность кусочков текста на разные темы.

Если я правильно поняла, то это в некотором смысле аналог мультимодальной модели, описанной выше. Интересно было бы посмотреть на их сравнение. Про плавность смены тем: а не решают ли эту проблему какими-то прямыми методами моделирования зависимости между метками соседних слов, например, с помощью скрытых марковских моделей?