Тематическое моделирование коротких текстовых сообщений

Введение

Тематическое моделирование — это метод обучения без учителя, способный сканировать набор документов, обнаруживать в них шаблоны слов и фраз и автоматически определять, к каким темам относится каждый из документов. В тематических моделях *из документов выделяются группы слов, интерпретация которых как тем должна помочь нам в лучшем понимании данных*.
Для тематического моделирования в основном используются два подхода: Латентное распределение Дирихле и Неотрицательная матричная факторизация. Однако традиционные методы тематического моделирования зависят от конкретного языка и полагаются на фиксированный словарный запас, специфический для языка обучения. Поэтому существующие модели сталкиваются с двумя проблемами: они не могут обрабатывать незнакомые слова и они не могут переноситься на другие языки.
В данной работе исследовались методы из семейства Contextualized Topic Models, которые внедряют контекстные векторные представления в нейронные тематические модели и демонстрируют более качественное разбиение на темы в сравнении с обычными тематическими моделями.

Цели и задачи практики

Целью данной работы было изучение и реализация одного из современных методов тематического моделирования: Contextualized Topic Models. В работе сделан акцент на изучение теории, которая стоит за алгоритмами для создания контекстных эмбеддингов, а также моделей для тематического моделирования при помощи таких векторных представлений.
Предложенные модели в дальнейшем использовались для кластеризации датасета научных статей.

Изученные материалы

1. Contextual Embeddings

Традиционно при обучении моделей для решения задач обработки естественного языка использовались глобальные векторные представления слов, такие как Word2Vec и Glove. Такие методы сначала строят словарь уникальных слов из всех документов, а затем создают для каждого слова векторное представление, учитывая при этом частоту его вхождения в документ. Но у таких методов есть несколько существенных недостатков:

- векторное представление для конкретного слова всегда фиксированное, вне зависимости от контекста, в котором оно встречается
- в таком статическом векторном представлении мы также не учитываем особенности семантики, синтаксиса, констатаций и т. д.

В отличие от традиционных представлений, контекстные эмбеддинги выходят за пределы словесного уровня анализа и каждому токену сопоставляют представление, которое является функцией всей входной последовательности. Такие контекстно-зависимые представления могут захватить много синтаксических и семантических свойств слов в различных лингвистических контекстах. Таким образом представление слова зависит от других слов в этом предложении, т к оно получается динамически из предобученной модели-трансформера (например BERT или ELMO), в которой механизм внимания (attention) смотрит на связь слова с его соседями.

В обеих моделях из семейства CTM (Contextualized Topic Models) предложено получать векторные представления из SBERT (Sentence-BERT) - модификации предобученной сети BERT, которая основана на двойной архитектуре (содержит два идентичных BERT с одинаковыми весами) и обрабатывает параллельно два предложения на этапе обучения, что позволяет быстро создавать семантически значимые представления *предложений*.

2. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence

Variational Autoencoder (VAE) - архитектура, состоящая из энкодера и декодера, которая обучается минимизации ошибки восстановления (reconstruction loss) между зашифрованными и исходными данными. При этом энкодер выдаёт не один вектор размера n, а два вектора размера n – вектор средних значений μ и вектор стандартных отклонений σ, что позволяет строить непрерывное скрытое пространство и выполнять случайные преобразования и интерполяцию.

Авторы статьи ставят перед собой цель научить модель создавать более осмысленные и согласованные темы заменяя традиционные тематические модели на объединение контекстуальных представлений с нейронными тематическими моделями.

1) *Идея:* На вход нейронной тематической модели (главное, чтобы она являлась расширением автоэнкодера) мы подаем предобученные контекстуальные эмбеддинги (главное, чтобы они представляли документы) и BoW представления.

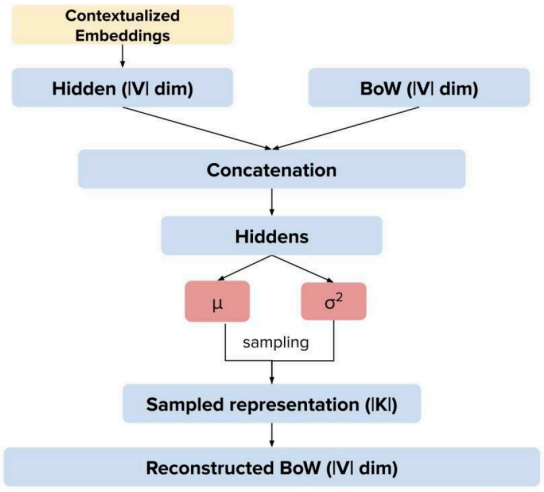

Figure 1: High-level sketch of CombinedTM.

2) Архитектура: В качестве нейронной тематической модели используется ProdLDA основанная на VAE, а векторные представления документа получают из SBERT.

Векторные представления документа проходят через скрытый слой и конкатенируются с BoW представлением.

Модель работает следующим образом: сначала она обучается напрямую отображать BoW представление документа в непрерывное скрытое пространство. Затем декодер реконструирует BoW, генерируя его слова из скрытого представления документа.

3. Cross-lingual Contextualized Topic Models with Zero-shot Learning

Zero-shot Learning — это постановка задачи в машинном обучении, когда во время теста модель наблюдает за выборкой из классов, которых не было во время обучения, и ему необходимо предсказать класс, к которому они принадлежат.

Авторы статьи постарались обойти два ограничения традиционных алгоритмов тематического моделирования: неумение работать с ранее незнакомыми словами и отсутсвие возможности применения той же модели к другим языкам.

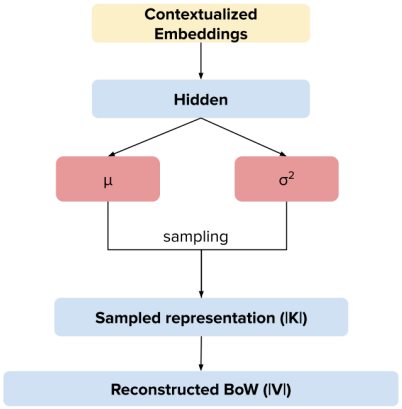

Figure 1: High-level schema of the architecture for the proposed contextualized neural topic model.

1) Идея: Модель строится в предположении, что суть всех тем одинакова для разных языков, а значит можно выучить смысл на одном языке и затем применять к другим языкам. При этом для обучения используется zero-shot learning: мы обучаем модель на одном языке, а тестируем на языках, которые она не видела на этапе обучения. И в качестве входных данных обязательно используются контекстуальные эмбеддинги, которые также являются многоязычными, для обеспечения корректной работы алгоритма. Таким образом авторы статьи предлагают новую нейронную модель, в которой BoW (мешок слов) заменяется многоязычными контекстуальными векторными представлениями. Такие представления мы получаем из соответствующей предобученной нейронной модели и подаем их на вход. Это позволяет нам работать с новыми словами на этапе теста и определять темы новых документов на языках, отличных от тех, что использовались в данных при обучении.

2) Архитектура: Авторы предлагают усовершенствовать модель Neural-ProdLDA, которая основана на Variational AutoEncoder и представляет собой нейронную модель для тематического моделирования. А именно они заменяют

входные BoW на предобученные представления из SBERT. Такие входные эмбеддинги учитывают контекст и порядок слов, а также дают возможность работать с другими языками.
При этом в архитектуре сохраняется финальный слой реконструированного BoW, так как он все еще необходим на этапе обучения для получения индикаторов тем (к примеру слов, которые с наибольшей вероятностью характеризуют тему).

Полученные результаты

В практической части своей работы я применяла модели CombinedTM и ZeroShotTM для кластеризации датасета научных статей.
Также передо мной стояла цель написать туториал о том, как работать с этими моделями и применять их на практике, потому что часто это не является чем-то очевидным и даже чтобы просто запустить модель нужно долго разбираться в исходном коде и статьях, а потом еще столько же разбираться, что данная модель выдает и как интерпретировать полученные результаты.

Подробное описание моей работы, пошаговая работа с моделями и полученные результаты можно найти **здесь**.

Заключение

После изучения моделей стало понятно, что CombinedTM комбинирует контекстуальные представления с традиционным пакетом слов, за счет чего делает более последовательные и согласованные темы, а ZeroShotTM является идеальной моделью для задач, в которых мы могли бы иметь недостающие слова в тестовых данных или хотели бы без особых вычислительных сложностей использовать модель для языков, отличных от языка обучения. При этом обе модели хорошо справились с кластеризацией датасета научных статей на английском языке.

Календарный план-график

№ п/п	Сроки проведения	Выполняемые работы
1	01.07.2022	Инструктаж по ознакомлению с требованиями охраны труда, техники безопасности, пожарной безопасности, а также правилами внутреннего трудового распорядка
2	02.07.2022 - 05.07.2022	Изучение Contextualized Topic Modeling на основе статей
3	06.07.2022 - 10.07.2022	Написание модели кластеризации для датасета научных статей
4	11.07.2022 -14.07.2022	Написание туториал по использованию Contextualized Topic Modeling.
5	15.07.2022	Обсуждение результатов с руководителем практики

Список источников

- Cross-lingual Contextualized Topic Models with Zero-shot Learning, https://aclanthology.org/2021.eacl-main.143.pdf
- Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, https://aclanthology.org/2021.acl-short.96.pdf
- Contextualized Topic Models, https://github.com/MilaNLProc/contextualized-topic-models
- Sentence Transformers, SBERT, https://www.pinecone.io/learn/sentence-embeddings/