Автоматическая аннотация тем в вероятностном тематическом моделировании

A Preprint

Карпинская Анна Викторовна
Московский государственный университет имени М. В. Ломоносова
Научный руководитель: Воронцов Константин Вячеславович
annakarpinckaya@gmail.com

Abstract

Вероятностное тематическое моделирование широко используется для выявления скрытых тем в текстах, что делает его важным инструментом для анализа больших текстовых корпусов. Задача именования тем подразумевают понимание смысла этих тем и создание понятных названий для каждой из них. В данной работе рассматриваются методы автоматического именования и суммаризации тем, выделенных в процессе вероятностного тематического моделирования. Приводятся подходы, основанные на оценке релевантности слов для тем, а также методы генерации кратких описаний для облегчения их интерпретации. Проводится анализ эффективности предложенных методов на текстовых корпусах разной тематики, что подтверждает их значимость для применения результатов тематического моделирования.

Keywords Тематическое моделирование · Автоматическая аннотация · Суммаризация текста · Генеративные модели · Метрики качества текста · Вероятностные модели · LDA · T5 · NLP · Семантическая близость

1 Введение

Современные технологии обработки текста и анализа данных способствовали развитию вероятностного тематического моделирования, которое используется для выявления скрытых тем в больших текстовых коллекциях. [1] Это направление нашло широкое применение в задачах анализа документов, информационного поиска и построения автоматических рекомендаций. Однако использование результатов тематического моделирования ограничивается сложностью их интерпретации, так как модели часто предоставляют лишь наборы ключевых слов, которые не всегда позволяют понять суть выявленных тем. [2]

Для решения этой проблемы активно исследуются методы автоматической аннотации и суммаризации тем. Такие подходы позволяют создавать понятные текстовые описания, которые упрощают интерпретацию и применение результатов моделирования. Одним из ключевых направлений является интеграция вероятностных моделей с генеративными языковыми моделями, которые обеспечивают связность и читаемость итоговых текстов.

Особенно актуальной задачей является адаптация систем аннотации к коллекциям документов разной тематики и объема. Например, использование современных моделей, таких как T5 и GPT, в сочетании с тематическими моделями, позволяет формировать краткие описания тем на основе ключевых фраз и репрезентативных фрагментов текста. Эти методы способны не только повышать точность, но и улучшать восприятие результатов тематического анализа. [3]

В данной статье предлагается методика, которая объединяет вероятностное тематическое моделирование, методы извлечения ключевых фраз и генеративные модели. Представленные подходы направлены на повышение связности и релевантности генерируемых текстов. Для оценки эффективности подхода

используются стандартные метрики, такие как ROUGE, BLEU и METEOR, а также косинусное сходство векторных представлений текстов. Экспериментальная часть включает тестирование на реальных данных, что позволяет оценить общее качество генерируемых аннотаций.

2 Постановка задачи

Задача автоматической аннотации и суммаризации тем в текстовых коллекциях представляет собой задачу генерации кратких и содержательных описаний для набора документов, объединённых в темы. [langston2024automated] В данном исследовании тексты новостных статей из набора данных BBC News используются для создания таких описаний.

В данном случае тексты статей представляют собой документированные данные, разбитые по темам.

Множество документов $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, где каждый документ D_i описан текстом $T(D_i)$.

Множество тем $\mathcal{T} = \{T_1, T_2, \ldots, T_K\}$, где каждая тема T_k содержит подмножество документов $\mathcal{D}_k \subseteq \mathcal{D}$.

Каждая тема содержит набор статей, которые необходимо обработать для извлечения репрезентативных фрагментов и ключевых фраз. [4] На основе этих данных требуется сгенерировать связное описание, которое бы отражало основное содержание темы.

Выходные данные: Для каждой темы T_k необходимо создать описание S_k, которое удовлетворяет следующим критериям:

- Является связным текстом.
- Содержит ключевые аспекты документов, входящих в $\mathcal{D}_k$.
- Является кратким, но информативным.

Задача:

- Извлечь репрезентативные фрагменты из текстов документов $\mathcal{D}_k$, чтобы охарактеризовать основное содержание темы.
- Формирование набор ключевых фраз, отражающих основные концепты темы.
- Генерация связное описание S_k, используя комбинацию репрезентативных фрагментов и ключевых фраз с помощью генеративной модели (например, T5).

Таким образом, итоговая задача сводится к построению функции $G(T_k)$, которая для темы T_k из множества документов $\mathcal{D}_k$ создаёт текст S_k:

$$G(T_k) = S_k, \quad \text{где } S_k = \text{генерируемое описание темы.}$$

3 Описание алгоритма

Для автоматической аннотации и суммаризации тем в текстах используется алгоритм, который включает в себя три основных этапа:

- Отбор репрезентативных фрагментов
- Извлечение ключевых фраз
- Генерация текстовых описаний

Каждый этап реализуется с применением современных языковых моделей и методов тематического моделирования. В качестве основы используются вероятностное тематическое моделирование и генеративные модели, такие как T5.[5]

3.1 Извлечение репрезентативных фрагментов

На первом этапе документы, входящие в одну тему, объединяются в общий текст. Для каждого предложения в объединённом тексте рассчитывается оценка релевантности теме, основанная на частотности ключевых токенов. Репрезентативные фрагменты определяются как предложения с наивысшими значениями релевантности:

- Все тексты проходят лемматизацию и удаление стоп-слов для формирования множества токенов.
- Для каждого предложения вычисляется его релевантность как сумма частот ключевых токенов, встречающихся в нём.
- Выбираются N_{top} предложений с максимальными значениями релевантности.

3.2 Извлечение ключевых фраз

Для извлечения ключевых фраз используются современные модели эмбеддингов, такие как Sentence-BERT [6]:

- Кандидатные ключевые фразы извлекаются из объединённого текста с использованием методов выделения именных групп.
- Для каждого документа и ключевой фразы вычисляется эмбеддинг. Центроидный вектор темы формируется как среднее значение векторов всех документов.
- Косинусное сходство между векторами ключевых фраз и центроидным вектором темы используется для выбора наиболее релевантных ключевых фраз.

3.3 Генерация описаний

Используя репрезентативные фрагменты и ключевые фразы, формируется входной промпт для генеративной модели:

- Входные данные включают список ключевых фраз и репрезентативных предложений.
- Генеративная модель (в данном случае T5) генерирует текстовое описание темы, основываясь на предоставленном промпте.[7]

3.4 Оценка качества

Для оценки качества генерируемых описаний используются метрики, которые измеряют релевантность, связность и полноту текстов. Основные используемые метрики:

ROUGE — измеряет пересечение n-грамм между сгенерированным текстом и эталоном:

$$P = \frac{\text{Совпадающие } n\text{-граммы}}{\text{Общее количество } n\text{-грамм в предсказании}}$$

$$R = \frac{\text{Совпадающие } n\text{-граммы}}{\text{Общее количество } n\text{-грамм в эталоне}}$$

$$F_1 = \frac{2PR}{P + R}$$

BLEU — использует n-граммную точность, взвешенную с учётом штрафа за длину:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

METEOR — учитывает лемматизацию, синонимы и порядок слов:

$$\text{METEOR} = 10 \cdot \frac{\text{Совпадения}}{\text{Средняя длина текста}}$$

Семантическая близость — вычисляется через косинусное сходство между эмбеддингами сгенерированного текста и эталона:

$$\text{Sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

Эти метрики позволяют объективно оценить качество аннотации и выявить сильные и слабые стороны алгоритма.

4 Эксперименты

4.1 Описание данных

Для экспериментов использовались данные из открытого набора BBC News Data, который включает текстовые данные пяти категорий: business, entertainment, politics, sport и tech. Каждый документ представляет собой новостную статью, сгруппированную по теме.

Для каждой категории предоставлены наборы текстов, которые позволяют провести тематическое моделирование и автоматическую аннотацию.

В Таблице 1 ниже представлено распределение данных по категориям:

Категория	Количество документов	Пример содержания
Business	510	Экономика, финансы, корпоративные новости
Entertainment	386	Кино, музыка, шоу-бизнес
Politics	417	Политические события, выборы
Sport	511	Новости спорта, результаты матчей
Tech	401	Технологические разработки, гаджеты

Таблица 1: Статистика категорий корпуса данных

4.2 Предварительно обученные модели NLP

Для реализации алгоритма аннотации и суммаризации тем использовались предварительно обученные языковые модели, способные эффективно обрабатывать текст на английском языке. Эти модели обучены на больших корпусах данных и способны извлекать как контекстуальные, так и семантические зависимости между словами.

В данном исследовании основное внимание уделяется применению модели T5 (Text-to-Text Transfer Transformer) для генерации описаний тем. [8] T5 адаптирована для задач как суммаризации текста, так и обработки контекстных данных.

4.3 Архитектура модели T5

Модель T5 представляет собой мощную архитектуру Transformer, которая переводит задачи NLP в единую текстовую форму. На этапе экспериментов использовалась версия модели, предварительно обученная на корпусе C4 (Colossal Clean Crawled Corpus). [4] Эта архитектура позволяет:

- Генерировать связные описания тем.
- Учитывать широкий контекст текстовых данных.
- Эффективно адаптироваться к задачам аннотации.

4.4 Обоснование выбора модели

Для данного исследования T5 была выбрана по следующим причинам:

1. Гибкость: модель способна адаптироваться к различным задачам, включая тематическую суммаризацию и генерацию текстов.
2. Качество генерации текста: демонстрирует высокие показатели на эталонных наборах данных для задач суммаризации.
3. Поддержка сложных структур: позволяет учитывать как длинные, так и короткие текстовые контексты.

Для получения эмбеддингов, используемых в оценке семантической близости, применялись Sentence-BERT и Universal Sentence Encoder. Эти модели помогают оценить качество сгенерированных текстов по отношению к исходным.

4.5 Методы оценки

Для оценки качества аннотаций и суммаризации тем использовались следующие метрики [9]:

- ROUGE: для измерения пересечения n-грамм между сгенерированным текстом и эталоном.
- BLEU: для оценки точности генерации на уровне n-грамм.
- METEOR: учитывает синонимы, порядок слов и точность в тексте.
- Семантическая близость: косинусное сходство эмбеддингов текста.

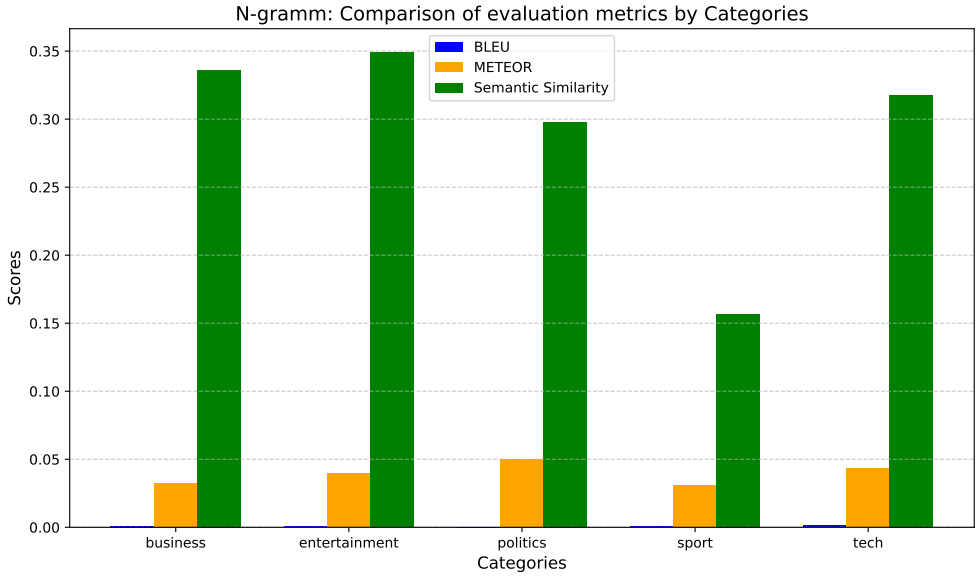

Рис. 1: Сравнение метрик BLEU, METEOR и Semantic Similarity

По результатам анализа на Рис.1 можно заметить, что семантическая близость демонстрирует значительно более высокие значения по сравнению с метриками BLEU и METEOR. Это указывает на способность алгоритма улавливать контекст и создавать описания, которые близки по смыслу к исходным данным, даже при наличии отклонений в синтаксической структуре. С другой стороны, BLEU-метрика показывает низкие значения для всех категорий, что говорит о различиях между сгенерированными описаниями и эталонами в последовательности слов и точном совпадении n-грамм. Модель создает более вариативные и свободные описания, которые не всегда точно соответствуют тексту эталона.

На Рис.2 видно, что ROUGE-1 метрика демонстрирует наивысшие значения для категорий politics и tech, что свидетельствует о хорошей способности модели передавать ключевые слова и выражения из исходных текстов. Однако ROUGE-2 метрика показывает значительно более низкие значения по сравнению с ROUGE-1, что говорит о том, что модель имеет затруднения с передачей более длинных последовательностей слов.

5 Заключение

Задача автоматической аннотации и суммаризации тем была формализована как задача генерации краткого и связного текста, который описывает основное содержание документов, принадлежащих к одной теме.

В рамках исследования использовался подход, основанный на вероятностном тематическом моделировании и генеративных моделях, таких как T5, для формирования текстовых описаний. Для оценки качества сгенерированных текстов применялись стандартные метрики: ROUGE, BLEU, METEOR и семантическая близость. Эти метрики позволили получить представление о соответствии сгенерированных текстов исходным данным.

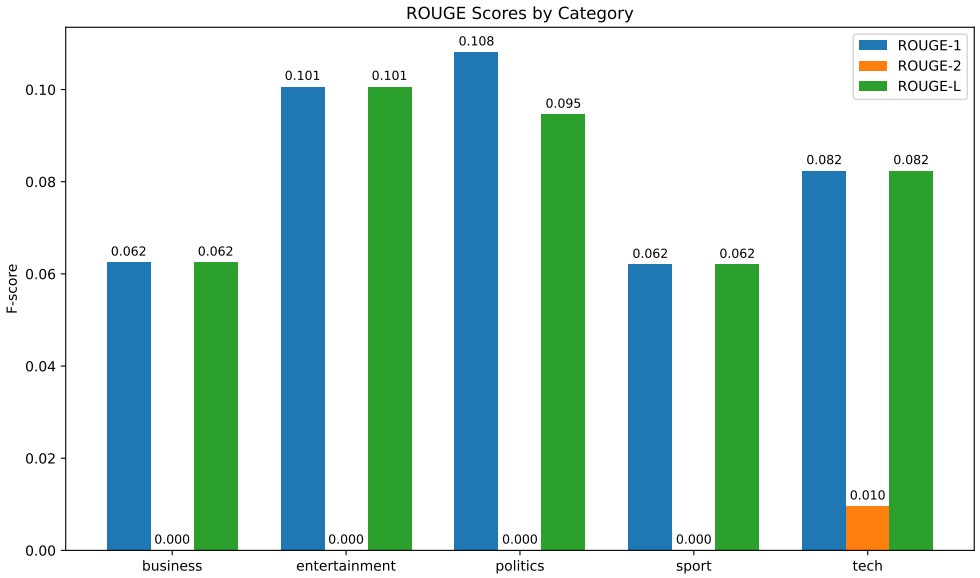

Рис. 2: Сравнение ROUGE-метрик

Результаты экспериментов показали, что метрика семантической близости демонстрирует наиболее высокие значения по сравнению с BLEU и METEOR. Это свидетельствует о способности модели учитывать контекст текста и создавать описания, соответствующие исходному содержанию.

ROUGE-метрики варьируются в зависимости от категории текста, что показывает различия в сложности генерации кратких описаний для каждой темы. Например, наилучшие результаты наблюдаются для категорий politics и tech, что может быть связано с высокой структурированностью данных в этих категориях.

Обзор литературных источников и результаты экспериментов подтверждают, что использование современных моделей, таких как T5, в сочетании с методами оценки качества, основанными на семантической близости, позволяет значительно улучшить результаты автоматической аннотации.

Список литературы

1. Blei D. M. Введение в вероятностные тематические модели // Columbia University. — 2011. — URL: https://www.cs.columbia.edu/~blei/papers/Blei2011.pdf.
2. Perez M., Nguyen A. Понятные тематические модели для кластеризации документов // arXiv preprint. — 2020. — URL: https://arxiv.org/pdf/1309.6874.
3. Automatic Labelling of Topic Models / J. H. Lau [и др.] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — 2011. — URL: https://www.researchgate.net/publication/220874747_Automatic_Labelling_of_Topic_Models.
4. Langston O., Ashford B. Автоматизированное обобщение тезисов и содержания нескольких документов с использованием больших языковых моделей // TechRxiv. — 2024. — URL: https://www.techrxiv.org/doi/full/10.36227/techrxiv.172262754.45577350.
5. Rönnqvist S. Исследовательское тематическое моделирование с использованием дистрибутивной семантики // arXiv preprint. — 2015. — URL: https://arxiv.org/pdf/1507.04798.
6. Киселев М. В., Пивоваров В. С., Шмулевич М. М. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов... // Информационные Материалы конференции. — 2005. — URL: https://elar.urfu.ru/bitstream/10995/1421/1/IMAT_2005_22.pdf.
7. Daumé III H., Marcu D. Настройка выравнивания слов и фраз для автоматического суммирования документов // arXiv preprint. — 2009. — URL: https://arxiv.org/pdf/0907.0804.
8. Attention is all you need / A. Vaswani [и др.] // Advances in Neural Information Processing Systems. — 2017. — С. 5998—6008.

9. BLEU: a method for automatic evaluation of machine translation / K. Papineni [и др.] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. — ACL. 2002. — С. 311—318.