Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем

Щукин Илья Вячеславович

Реализация алгоритма TextTiling для решения задачи сегментации диалогов

Отчёт по учебной практике

Научный руководитель:
ассистент кафедры ИАС Чернышев Г. А.

Санкт-Петербург
2024

Оглавление

1 Введение

Семантическая сегментация диалогов является важной задачей из области моделирования диалогов. Задачу можно сформулировать следующим образом: необходимо для данного текста в виде диалога найти такое разбиение на сегменты, чтобы сообщения внутри каждого сегмента были объединены одной темой обсуждения.

В данной работе описывается разработка и модификации алгоритма TextTiling для решения задачи сегментации диалогов. Эффективное решение данной задачи в дальнейшем позволит реализовать систему для выделения ключевых тем из диалогов (суммаризации). Подобная система может быть использована как дополнение к существующим социальным сетям и мессенджерам, так и к корпоративным системам для поддержания баз знаний.

2 Постановка задачи

Целью данной работы является реализация алгоритма TextTiling. Для её достижения были сформулированы следующие задачи:

- Выполнить обзор предметной области.
- Реализовать алгоритм TextTiling.
- Экспериментально сравнить модели для оценки когерентности.

3 Обзор

3.1 Алгоритм TextTiling

В статье [3] приводится алгоритм состоящий из трёх шагов:

1. токенизация,
2. получение оценок когерентности,
3. определение границ.

Токенизация На первом этапе входной текст разбивается на отдельные лексические единицы. Также на данном этапе производится нормализация текста, т.е. он приводится к единому формату: весь текст переводится в нижний регистр и из него удаляются слова из списка стоп слов.

Оценка когерентности Для оценки когерентности авторы предлагают два подхода: блочный, где для прилегающих блоков текста подсчитывается число совпадающих токенов и метод подсчета новых токенов, в котором для получения оценок для отрезков подсчитывается число ранее не встречавшихся токенов.

Определение границ Границы определяются с помощью оценок глубины. Оценки глубины вычисляются на основе оценок когерентности по следующей формуле:

$$depth(i) = score(l) + score(r) - 2score(i)$$

Где $score(l)$ и $score(r)$ это наибольшие оценки когерентности левее и правее пробела i между сообщениями. Далее эти оценки упорядочиваются. Большая глубина указывает на то, что в этом месте должна располагаться граница диалога. Поэтому значения больше определенного порога отсекаются, и в них помещаются границы.

3.2 Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring

Для решения задачи сегментации авторы данной работы [9] используют алгоритм TextTiling, где модель BERT [1] используется для получения оценок когерентности сообщений.

BERT — это энкодерная модель на основе архитектуры transformer. Одной из задач при обучении BERT была задача next sentence prediction. В данной задаче модели требуется предсказать идут ли два предложения последовательно или нет. Поэтому эту модель можно использовать на этапе получения оценок когерентности. Модель без дообучения способна показывать метрики значительно превосходящие метод случайной сегментации, но исходно она обучалась на данных из Wikipedia и корпусе книг, поэтому авторы данной статьи дообучают её на другом домене.

3.3 SentenceTransformers

SentenceTransformers — это фреимворк с множеством предобученных моделей для задачи получения эмбеддингов из текстов, изображений [7]. В данной работе используется модель all-mpnet-base, которая в свою очередь является дообученным вариантом модели mpnet [4].

3.4 SuperDialseg

В данной публикации [8] авторы создают собственный синтетический датасет на основе датасетов Doc2Dial [10] и MultiDoc2Dial [5]. В данных датасетах содержатся диалоги, информация в которых подкрепляется отдельными документами. Для каждого сообщения есть ссылка на какую-то часть соответствующего документа. Подобные наборы данных могут быть использованы например для выравнивания больших языковых моделей, поскольку подкрепление диалогов реальными данными уменьшает галлюцинирование модели после выравнивания. Данные диалоги размечались вручную, что делает эти данные

более чистыми, чем получаемые с помощью скраппинга из интернета.

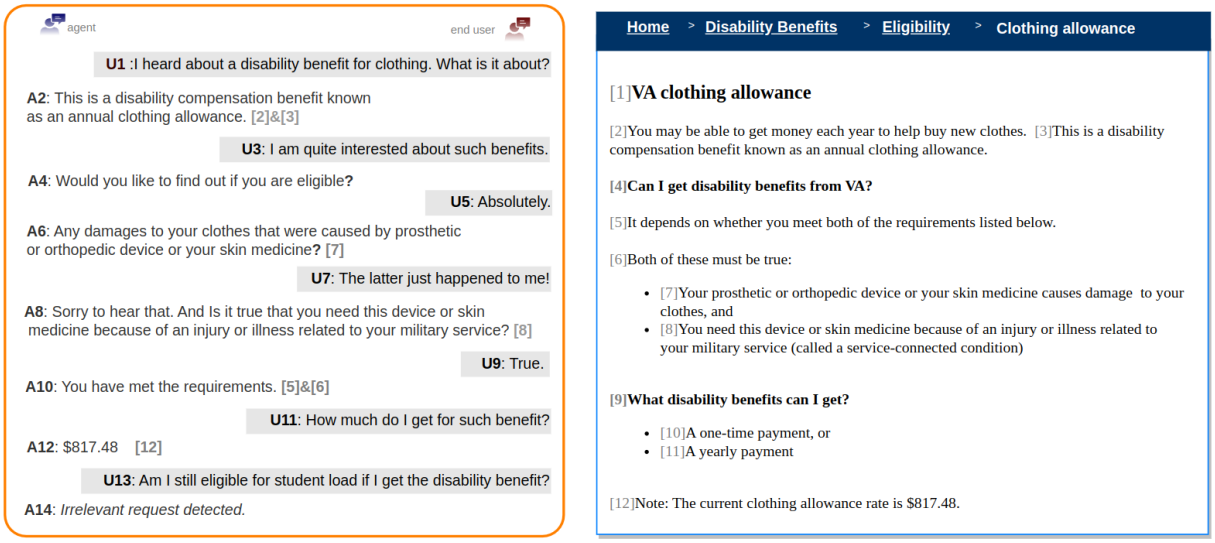

Рис. 1: Пример диалога из Doc2Dial. Для каждого сообщения есть соответствующая ссылка (в квадратных скобках)[1]

Для создания SuperDialseg авторы соотносят части диалогов с частями документов, на которые они ссылаются. Если часть диалога ссылается на одну и ту же часть документа, то это означает, что в ней обсуждается одна тема, а смена ссылки на другую часть документа обозначает смену темы обсуждения. Данный метод позволяет создать необходимую для задачи сегментации разметку.

В данной работе датасет SuperDialseg используется для тестирования решения, а также для дообучения модели BERT.

[1]Изображение с сайта https://doc2dial.github.io/

4 Реализация

4.1 Алгоритм TextTiling

При реализации были использованы следущие средства и библиотеки:

- Python3.9
- Hugging Face transformers
- NLTK
- SentenceTransformers

В исходном описании алгоритма этап токенизации не зависит от этапа получения оценок когерентности, но в реализации полученной в данной работе это не так, так как разные модели требуют разной токенизации.

4.2 Методы оценки когерентности

4.2.1 Случайный

Случайный метод равномерно выдает значения в пределах от 0 до 1 для каждого пробела между сообщениями. Если значение больше 0.5, то в этом пробеле устанавливается граница. Случайный метод сегментации необходим для сравнения эффективности алгоритмов. Получаемые метрики для случайного метода показывают верхнюю границу для метрик качества.

4.2.2 BERT

В полученной реализации используется вариант BERT-base с 110 млн. параметров. Для получения оценки с помощью BERT модели на вход подается пара сообщений, разделенная специальным токеном [SEP]. К получаемым логитам затем применяется SoftMax, который показывает вероятности для двух классов “последовательные”, “непоследова-

тельные". Получаемая вероятность затем используется в качестве оценки когерентности.

4.2.3 Sentence-transformers

Для каждого предложения получается его представление в виде вектора — эмбеддинга. Оценка когерентности в данном случае равна косинусному расстоянию между получаемыми эмбеддингами.

4.2.4 Дообучение BERT

Также была подготовлена дообученная версия BERT на сообщениях из датасета. Для этого была выбрана часть диалогов, которые не попали в тестирующую выборку. Далее по этим диалогам был построен датасет из троек сообщений, где первые два идут последовательно в тексте, а третье берется случайным образом, при этом гарантируется что третье сообщение не совпадает с двумя предыдущими. Такой датасет необходим для возможности использовать margin ranking loss:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} max(0, \eta + c_i^- - c_i^+),$$

где N — размер датасета, η — гиперпараметр, а c_i^+ — оценка когерентности для пары последовательных предложений, а c_i^- — для пары непоследовательных.

5 Эксперименты

5.1 Оценка качества

Для оценки качества сегментации используются метрики WindowDiff и Pk.

WindowDiff [6] вычисляется по следующей формуле:

$$WindowDiff(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0),$$

где ref — исходное разбиение диалога, hyp — предсказанное разбиение, k — размер окна, N — размер диалога, а $b(i, j)$ — число границ между позициями i и j. В экспериментах размер окна равен 3.

Pk [2] вычисляется следующим образом:

$$D_\mu(i, j) = \gamma_\mu e^{-\mu|i-j|},$$

$$P_\mu(ref, hyp) = \sum_{1\le i\le j\le n} D_\mu(i, j)\delta_{ref}(i, j)\overline{\oplus}\delta_{hyp}(i, j),$$

где σ — индикаторная функция равная одному, если два сообщения принадлежат одному сегменту в разбиении, $\overline{\oplus}$ — XNOR (“оба или никакой”). При этом в качестве параметра μ выбирается половина средней длины сегмента в истинном разбиении.

5.2 Результаты

Таблица 1: Сравнение метрик для разных методов получения оценок.

Модель	WindowDiff ↓	Pk ↓
Random	0.613	0.504
BERT	0.557	0.463
BERT finetuned	0.541	0.428
Sentence Transformers all-mpnet	**0.527**	**0.414**

В Таблице 1 представлены метрики полученные для разных моделей на тестовом датасете. Видно, что полученные реализации показывают

лучшее качество, чем случайный метод. Также видно, что дообучение модели BERT улучшает итоговое качество предсказаний. Лучше всего себя показывает метод использующий модель all-mpnet в качестве энкодера сообщений.

6 Заключение

В ходе работы были достигнуты следующие результаты:

1. Выполнен обзор предметной области.
2. Реализован алгоритм TextTiling.
3. Проведено сравнение моделей для оценки когерентности.

С кодом проекта можно ознакомиться на github[2].

[2]https://github.com/Elluran/text-tiling

Список литературы

[1] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // CoRR. — 2018. — Vol. abs/1810.04805. — arXiv : 1810.04805.

[2] Beeferman Doug, Berger Adam, and Lafferty John. Text Segmentation Using Exponential Models // Second Conference on Empirical Methods in Natural Language Processing. — 1997. — Access mode: https://aclanthology.org/W97-0304.

[3] Hearst Marti A. Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages // Computational Linguistics. — 1997. — Vol. 23, no. 1. — P. 33–64. — Access mode: https://aclanthology.org/J97-1003.

[4] Song Kaitao, Tan Xu, Qin Tao, Lu Jianfeng, and Liu Tie-Yan. MPNet: Masked and Permuted Pre-training for Language Understanding. — 2020. — 2004.09297.

[5] Feng Song, Patel Siva Sankalp, Wan Hui, and Joshi Sachindra. MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing / ed. by Moens Marie-Francine, Huang Xuanjing, Specia Lucia, and Yih Scott Wen-tau. — Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. — 2021. — Nov. — P. 6162–6176. — Access mode: https://aclanthology.org/2021.emnlp-main.498.

[6] Pevzner Lev and Hearst Marti A. A Critique and Improvement of an Evaluation Metric for Text Segmentation // Computational Linguistics. — 2002. — Vol. 28, no. 1. — P. 19–36. — Access mode: https://aclanthology.org/J02-1002.

[7] Reimers Nils and Gurevych Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. — 2019. — 1908.10084.

[8] Jiang Junfeng, Dong Chengzhang, Kurohashi Sadao, and Aizawa Akiko. SuperDialseg: A Large-scale Dataset for Supervised Dialogue Segmentation. — 2023. — 2305.08371.

[9] Xing Linzi and Carenini Giuseppe. Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring // CoRR. — 2021. — Vol. abs/2106.06719. — arXiv : 2106.06719.

[10] Feng Song, Wan Hui, Gunasekara Chulaka, Patel Siva, Joshi Sachindra, and Lastras Luis. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / ed. by Webber Bonnie, Cohn Trevor, He Yulan, and Liu Yang. — Online : Association for Computational Linguistics. — 2020. — Nov. — P. 8118–8128. — Access mode: https://aclanthology.org/2020.emnlp-main.652.