Классификация фрагментов текста на основе их явных векторных представлений

Отчёт по НИР

Кузнецов Алексей Артемович[1] and Воронцов Константин Вячеславович[2]

[1]Аспирант 1-го курса кафедры интеллектуальных систем, МФТИ (НИУ)
[2]Научный руководитель, д.ф.-м.н., профессор РАН

18 декабря 2023 г.

1 Введение

Задача разметки текста является одной из ключевых в области обработки естественного языка, которая представляет собой задачу поиска и классификации фрагментов в тексте (спанов). Она может встречаться в различных постановках, наиболее распространённой из которых является NER [7] (или NEL, Named Entity Labeling), когда искомые фрагменты текста являются именами собственными: названиями мест и организаций, именами людей, упоминаниями чисел и времени и т. д. Также поиск фрагментов в тексте используется в таких задачах, как question answering, constituent labeling [10], semantic role labeling [8], mention detection [1] и др.

Существует большое количество различных методов разметки текста. Наиболее распространённым подходом является sequence labeling, который представляет из себя присвоению тега каждому отдельному токену из системы тегов, таких как BIO, IOB или IOBES [7]. MRC (Machine Reading Comprehension) модели представляют собой модели, которые направлены на обнаружение и классификацию текстовых фрагментов в заданном контексте [4]. Эти модели напрямую моделируют отношения между фрагментами, что повышает производительность в таких задачах, как распознавание именованных сущностей и ответы на вопросы. В данном подходе в модель передаётся текст для разметки и инструкция, также сформулированнуя в виде текста. Также с развитием генеративных моделей стало возможным размечать текст с помощью генерации нового текста, содержащего информацию о найденных фрагментах.

В данной работе будет рассмотрен подход классификации спанов на основе их эмбеддингов, а также методы получения этих эмбеддингов.

Цель работы. Оценить возможность поиска и классификации фрагментов текста различной длины на основе создания и классификации их явных векторных представлений, получаемых из контекстуализированных эмбеддингов токенов фрагментов текста. Сравнить различные методы построение эмбеддингов спанов.

2 Обзор литературы

Дополнить обзор литературы...

Эмбеддинг фрагмента текста должен содержать как можно больше информации о содержании данного фрагмента и о его контексте. По этой причине подавляющее большинство методов построено на основе конкатенации эмбеддингов токенов, расположенных на границе спана, или аггрегации эмбеддингов всех токенов спана. Одним из ранних подходов для создания векторных представлений фрагментов текста является использование LSTM и BiLSTM сетей. В [2] для решения задачи анализа структуры аргументации используется конкатенация скрытых состояний на границе фрагмента обоих слоёв и их суммы и разности двуслойной BiLSTM сети с добавлением дополнительной информации о длине спана в виде вектора.

Существуют различные методы получения векторных представлений фрагментов текста, которые используют контекстуализированные эмбеддинги токенов. Так, в [1] описано шесть различных методов получения эмбеддингов спанов на основе предобученных векторных представлениях токенов и проведено их сравнение для шести задач. В качестве аггрегирующей функции может быть применён механизм внимания, выход которого конкатенируется с эмбеддингами токенов начала и конца фрагмента [5; 6].

Векторное представление фрагмента текста может быть использовано не только для явной классификации, но и как часть модели трансформера: в модели DSpERT [11] используется Span Transformer Block, в который в качестве вектора query подаются векторные представления спанов. В [3] представлен классификатор фрагментов текста в виде надстройки над моделью SpanBERT [9]. Архитектура описанного классифкатора аналогична механизму самовнимания в трансформерах.

3 Основная часть

3.1 Постановка задачи

Обозначения, которые будут использованы в данной работе: $\mathcal{D} = \{w_i\}_{i=1}^{N}$ — текст длины N как последовательность токенов w_i, $\mathcal{S}_{i,j} = \{w_i, ..., w_j\}$ — фрагмент текста с i-го по j-ый токен, $s_{i,j}$ — эмбеддинг фрагмента $\mathcal{S}_{i,j}$, e_i — эмбеддинг i-го токена, $\mathcal{C}_k$ — класс, K — количество целевых классов, $[\ ;\ ]$ — опреация конкатенации.

В общем виде задача формулируется следующим образом. Дан текст $\mathcal{D}$, состоящий из токенов $w_1, w_2, ..., w_N$. Для каждой пары индексов $(i, j) = \{i, j \mid 0 \leqslant i \leqslant j \leqslant N\}$, необходимо найти класс $\mathcal{C}_k \in \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\} \cap \{None\}$, которому принадлежит фрагмент $\mathcal{S}_{i,j} = \{w_i, ..., w_j\} \subset \mathcal{D}$. В данном случае будет классифицирован каждый возможный фрагмент текста любой длины. Однако существенным ограничением является квадратичная сложность от длины текста. Некотороми из возможных обходов этого ограничения является лимитирование длины классифицируемого фрагмента или классификация фрагментов фиксированной длины с оптимизацией размера окна, с помощью которого выделяются фрагменты.

В настоящей работе для обучения классификатора будут использованы заранее созданные фрагменты текста их эмбеддинги.

3.2 Методы создания эмбеддингов фрагметов текста

В ходе исследования были рассмотрены подходы к созданию эмбеддингов спанов на основе векторных представлений токенов, которые изложены ниже.

Max pooling и average pooling. В max pooling каждая k-ая компонента выходного вектора равна максимальной k-ой компоненте среди входных векторов: $s^k_{i,j} = max(e^k_i, e^k_{i+1}, ..., e^k_j)$, в average pooling — усреднённой k-ой компоненте: $s^k_{i,j} = \frac{1}{j-i}\sum_{k=i}^{j} e^k_i$.

Attention pooling. Взвешенное среднее эмбеддингов e_i токенов из фрагмента текста:

$$\alpha_i = v \cdot e_i, \quad a_i = softmax(\alpha_i),$$

$$s_{i,j} = s^{attn}_{i,j} = \sum_{k=i}^{j} a_k \cdot e_k,$$

где v — обучаемый вектор весов. Развитием данного подхода является конкатенация результата прменения алгоритма attention pooling с эмбеддингами начального и конечного токена фрагмента текста и значения $\phi(\mathcal{S}_{i,j})$, отражающего длину спана:

$$s_{i,j} = [e_i; e_j; s^{attn}_{i,j}; \phi(\mathcal{S}_{i,j})]$$

Endpoint. Конкатенация эмбеддингов начала e_i и конца e_j фрагмента: $s_{i,j} = [e_i; e_j]$.

Diff-Sum. Конкатенация суммы и разности эмбеддингов начала и конца фрагмента: $s_{i,j} = [e_i + e_j; e_i - e_j]$.

Coherent. Эмбеддинги начального и конечного токенов фрагмента делится на четыре части: $e_i = [e^1_i; e^2_i; e^3_i; e^4_i]$, причём $e^1_i, e^2_i \in \mathbb{R}^a, e^3_i, e^4_i \in \mathbb{R}^b$, где a и b могут варьироваться. Эмбеддинг спана формируется следующим образом: $s_{i,j} = [e^1_i; e^2_j; e^3_i \cdot e^4_j]$.

Biaffine преобразование основано на BiLSTM. H_f и H_b — скрытые состояния на выходе двуслойной BiLSTM, к которым применяется следующее преобразование:

$$R^{sent} = H_f^{\mathrm{T}} U^1_{sent} H_b + (H_f + H_b)^{\mathrm{T}} U^2_{sent} + b,$$

где $U^1_{sent}, U^2_{sent}, b$ — обучаемые матрицы весов. Далее идет усреднение матрицы $R^{sent}_{n\times n}$:

$$v_i = \frac{1}{n}\sum_{j=1}^{n} R^{sent}_{ij} + w_i$$

и применяется max pooling:

$$s_{i,j} = MaxPool(v_i, ..., v_j).$$

3.3 Языковые модели

Языковые модели для получения контекстуализированных эмбеддингов токенов были выбраны следующие: BERT, RoBERTa, SpanBERT и XLNet. Анализ нескольких моделей необходим для получения более полного представления о методах создания эмбеддингов спанов и исследования вляния различных методов предобучения схожих по архитектуре моделей. Наибольший интерес из рассмотренных моделей представляет SpanBERT, поскольку при его обучении применялось маскирование не отдельных слов, а целых фрагментов, т. е. неразрывных

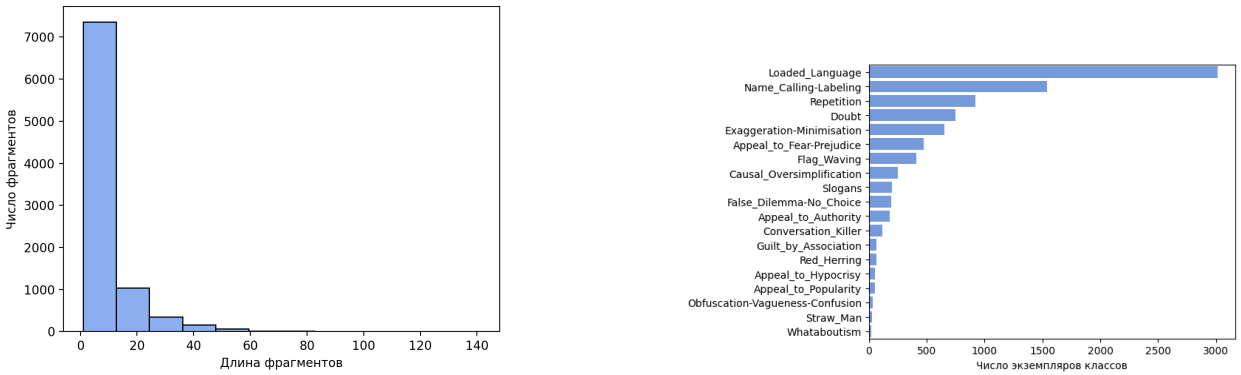

Рис. 1: Распределение длин фрагментов (слева) и распределение числа экземпляров по классам (справа).

подпоследовательностей токенов.

3.4 Датасет

При исследовании в качестве датасета были использованы данные, представленные на SemEval-2023. Датасет содержит 536 на английском языке, в каждой из которых размечены спаны с указанием пропагандистских техник, которые в них используются. Обучающая выборка состоит из 446 текстов с 7201 размеченным спаном, тестовая выборка — из 90 статей с 1801 размеченным спаном. Каждый спан в датасете принадлежит одному из 19 классов.

На рисунке 1 приведена статистика по датасету, отражающая распределение длин фрагментов и распределение количества классов.

Так как представленный в работе подход предполагает классификацию любого произвольного фрагмента текста, то для того, чтобы реализовать эту возможность, в датасет были добавлены случайные фрагменты текста, из которых был составлен класс *None*. Они были сэмплированы следующим образом: с помощью генератора псевдослучайных чисел генерировались индекс начала фрагмента из равномерного распределения и его длина из распределения, соответствующего распределению длин спанов в обучающей выборке. Число экзепляров класса *None* соответствует количеству экземляров двух самых многочисленных классов.

3.5 Классификатор

Дополнить описание классификатора...

Классификатор, использованный для предсказания класса фрагмента по его эмбеддингу, состоит из 3 полносвязных слоёв. Размерность скрытых состояний была выбрана 256. На выходе классификатора вектор с вероятностями принадлежности эмбеддинга каждому из 20 классов. Оптимизатор — AdamW, функция потерь — кросс-энтропия.

4 Результаты

Результаты сравнения классификации спанов по их эмбеддингам приведены в таблице. Сравнение производилось по метрике precision, усреднённой с весами по все классам.

Таблица 1: Взвешенное среднее точности.

	BERT	RoBERTa	SpanBERT	XLNet
Coherent	0,52	0,24	0,44	0,35
Diff-Sum	0,53	0,21	0,43	0,51
Endpoint	0,51	0,24	0,44	0,51
Avg pooling	0,46	0,31	0,35	0,44
Max pooling	0,34	0,26	0,32	0,40
Attention pooling	-	-	0,28	-

Из таблицы 1 видно, что лучший результат был показан при использовании метода diff-sum. Однако метод endpoint также дал высокие результаты. Это говорит о том, что построение эмбеддингов спанов на основе эмбеддингов крайних токенов фрагментов является наиболее перспективным. У аналогичных методов ткаже есть преимущество в виде отсутствия затрат на вычисление.

Ожидаемо низкие метрики классификации были получены из-за сильного дисбаланса классов. Точность классификации классов с наименьшим количеством экземпляров не превышала 0,05.

5 Заключение

В настоящей работе был исследован вопрос поиска и классификации фрагментов текста на основе их векторных представлений. Были проанализированы шесть методов построения эмбеддингов фрагметов на основе контекстуализированных эмбеддингов токенов, которые были созданы четырьмя языковыми моделями. Сравнение результатов классификации показало, что методы построения эмбеддингов спанов из эмбеддингов их крайних токенов работаю лучше всего. Помимо этого данные методы не требуют больших вычислительных ресурсов.

Хоть эксперименты не дали ожидаемого хорошего результата, тем не менее, было выяснено, что разметка текста на основе классификации эмбеддингов спанов является перспективным подходом, который будет исследован в дальнейшей работе.

Также в дальнейшую работу входит:

- повышение качества эксперимента: работа с дисбалансом классов, более тонкая настройка геперпараметров классификатора;
- расширение списка датасетов;
- разработка собсвенного метода создания эмбеддингов спанов и исследование влияния конкатенации эмбединга специального токена CLS в BERT-подобных моделях с эмбеддингом спана на точность классификации;
- исследование устойчивости методов разметки текста на основе классификации явных векторных представлений спанов к длине спанов;
- исследование методов разметки текста с использованием генеративных моделей как наиболее перспективных.

Список литературы

1. A Cross-Task Analysis of Text Span Representations / S. Toshniwal [и др.] // Workshop on Representation Learning for NLP. — 2020. — URL: https://api.semanticscholar.org/CorpusID:219531032.
2. An Empirical Study of Span Representation in Argumentation Structure Parsing / T. Kuribayashi [и др.] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — С. 4691—4698. — URL: https://www.aclweb.org/anthology/P19-1464.
3. CRSAtt: By Capturing Relational Span and Using Attention for Relation Classification / C. Shao [и др.] // Applied Sciences. — 2022. — Т. 12, № 21. — ISSN 2076-3417. — DOI: 10.3390/app122111068. — URL: https://www.mdpi.com/2076-3417/12/21/11068.
4. Dependency Parsing as MRC-based Span-Span Prediction / L. Gan [и др.] // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / под ред. S. Muresan, P. Nakov, A. Villavicencio. — Dublin, Ireland : Association for Computational Linguistics, 05.2022. — С. 2427—2437. — DOI: 10.18653/v1/2022.acl-long.173. — URL: https://aclanthology.org/2022.acl-long.173.
5. *Gandhi N.*, *Field A.*, *Tsvetkov Y.* Improving Span Representation for Domain-adapted Coreference Resolution // ArXiv. — 2021. — Т. abs/2109.09811. — URL: https://api.semanticscholar.org/CorpusID:237581144.
6. *Kahardipraja P.*, *Vyshnevska O.*, *Loáiciga S.* Exploring Span Representations in Neural Coreference Resolution // Proceedings of the First Workshop on Computational Approaches to Discourse / под ред. C. Braud [и др.]. — Online : Association for Computational Linguistics, 11.2020. — С. 32—41. — DOI: 10.18653/v1/2020.codi-1.4. — URL: https://aclanthology.org/2020.codi-1.4.
7. *Peres da Silva R.*, *Esteves D.*, *Maheshwari G.* Bidirectional LSTM with a Context Input Window for Named Entity Recognition in Tweets //. — 12.2017. — С. 1—4. — DOI: 10.1145/3148011.3154478.
8. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates / D. Larionov [и др.] // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) / под ред. R. Mitkov, G. Angelova. — Varna, Bulgaria : INCOMA Ltd., 09.2019. — С. 619—628. — DOI: 10.26615/978-954-452-056-4_073. — URL: https://aclanthology.org/R19-1073.
9. SpanBERT: Improving Pre-training by Representing and Predicting Spans / M. Joshi [и др.] // Transactions of the Association for Computational Linguistics. — 2019. — Т. 8. — С. 64—77. — URL: https://api.semanticscholar.org/CorpusID:198229624.
10. *Vilares D.*, *Gómez-Rodríguez C.* Discontinuous Constituent Parsing as Sequence Labeling // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) / под ред. B. Webber [и др.]. — Online : Association for Computational Linguistics, 11.2020. — С. 2771—2785. — DOI: 10.18653/v1/2020.emnlp-main.221. — URL: https://aclanthology.org/2020.emnlp-main.221.

11. *Zhu E.*, *Liu Y.*, *Li J.* Deep Span Representations for Named Entity Recognition. — 2023. — arXiv: 2210.04182 [cs.CL].