Классификация фрагментов текста на основе их явных векторных представлений

Научно-исследовательская работа

Кузнецов А. А.[1]
Научный руководитель: Воронцов К. В.

[1]Московский физико-технический институт
(национальный исследовательский университет)

December 18, 2023

Введение

Методы разметки текста:

- sequence labeing,
- MRC и генеративные модели,
- *классификация спанов на основе их 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.

Постановка задачи

Определение

Фрагмент текста (спан) — это непрерывная подпоследовательность токенов в тексте.

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 $\mathcal{S}_{i,j} = \{w_i, ..., w_j\} \subset \mathcal{D}$.

Подход к решению. Каждому возможному фрагменту $\mathcal{S}_{i,j}$ сопоставляется его эмбеддинг $s_{i,j}$, который классифицируется полносвязной нейронной сетью, состоящей из нескольких линейных слоев.

Рассматриваемые подходы к созданию эмбеддингов фрагментов

- Endpoint
- Diff-Sum
- Coherent
- Max pooling
- Average pooling
- Attention pooling

Toshniwal, Shubham et al. (2020) “A Cross-Task Analysis of Text Span Representations” https://arxiv.org/abs/2006.03866

Kahardipraja, Patrick, Olena Vyshnevska, and Sharid Loáiciga (2020) “Exploring Span Representations in Neural Coreference Resolution” https://aclanthology.org/2020.codi-1.4

Attention pooling

Attention pooling — взвешенное среднее эмбеддингов e_i токенов из фрагмента текста (v — обучаемый вектор параметров):

$$\alpha_i = v \cdot e_i, \quad a_i = softmax(\alpha_i), \quad s_{i,j} = s_{i,j}^{attn} = \sum_{k=i}^{j} a_k \cdot e_k$$

Развитие данного подхода:

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

Endpoint, Diff-Sum и Coherent

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:

$$s_{i,j} = [e_i + e_j; e_i - e_j].$$

В методе coherent эмбеддинги начального и конечного токенов фрагмента делится на четыре части: $e_i = [e_i^1; e_i^2; e_i^3; e_i^4]$, $e_i^1, e_i^2 \in \mathbb{R}^a, e_i^3, e_i^4 \in \mathbb{R}^b$. Эмбеддинг спана формируется следующим образом:

$$s_{i,j} = [e_i^1; e_j^2; e_i^3 \cdot e_j^4].$$

Языковые модели

- BERT
- RoBERTa
- SpanBERT
- XLNet

Devlin, Jacob et al. (2019) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”
https://api.semanticscholar.org/CorpusID:52967399

Liu, Yinhan et al. (2019) “RoBERTa: A Robustly Optimized BERT Pretraining Approach”
https://api.semanticscholar.org/CorpusID:198953378

Joshi, Mandar et al. (2019) “SpanBERT: Improving Pre-training by Representing and Predicting Spans”
https://api.semanticscholar.org/CorpusID:198229624

Yang, Zhilin et al. (2019) “XLNet: Generalized Autoregressive Pretraining for Language Understanding”
https://api.semanticscholar.org/CorpusID:195069387

Классификатор

Классификатор: полносвязная нейронная сеть из трёх линейных слоёв

Оптимизатор: AdamW

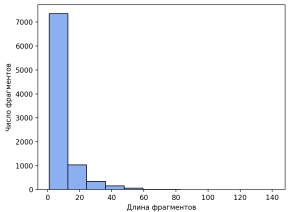
Функция потерь: $Loss = -\sum_{c=0}^{K} \hat{y}_c \log(p_c)$

Датасет SemEval 2023

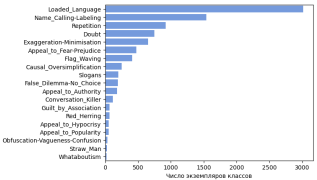
английский язык

Датасет состоит из 536 размеченных статей и содержит 9002 размеченных фрагментов на 19 классов.

	train	test
Число статей	446	90
Число фрагментов	7201	1801



В датасет были добавлены случайные фрагменты текста класса None: генерировались индекс начала фрагмента из равномерного распределения и его длина — из распределения длин спанов в обучающей выборке.



Da San Martino, Giovanni et al. (2019) “Fine-Grained Analysis of Propaganda in News Articles”

Результаты

Взвешенное среднее по метрике precision

	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	-

Заключение

Лучший результат был показан при использовании методов с использованием только крайних эмбеддингов фрагментов, которые вдобавок не требуют затрат на вычисление эмбеддингов спанов.

На данном этапе нельзя говорить универсальности рассмотренных подходов для поиска и классификации фрагментов текста произвольной длины.

Дальнейшая работа

- Повышение качества эксперимента: работа с дисбалансом классов, 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.

Ссылки I

- Da San Martino, Giovanni et al. (Nov. 2019). “Fine-Grained Analysis of Propaganda in News Articles”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. EMNLP-IJCNLP 2019. Hong Kong, China.
- Devlin, Jacob et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *North American Chapter of the Association for Computational Linguistics*. URL: https://api.semanticscholar.org/CorpusID:52967399.
- Joshi, Mandar et al. (2019). “SpanBERT: Improving Pre-training by Representing and Predicting Spans”. In: *Transactions of the Association for Computational Linguistics* 8, pp. 64–77. URL: https://api.semanticscholar.org/CorpusID:198229624.
- Kahardipraja, Patrick, Olena Vyshnevska, and Sharid Loáiciga (Nov. 2020). “Exploring Span Representations in Neural Coreference Resolution”. In: *Proceedings of the First Workshop on Computational Approaches to Discourse*. Online: Association for Computational Linguistics, pp. 32–41. DOI: 10.18653/v1/2020.codi-1.4. URL: https://aclanthology.org/2020.codi-1.4.

Ссылки II

- Liu, Yinhan et al. (2019). “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *ArXiv* abs/1907.11692. URL: https://api.semanticscholar.org/CorpusID:198953378.
- Toshniwal, Shubham et al. (2020). “A Cross-Task Analysis of Text Span Representations”. In: *CoRR* abs/2006.03866. arXiv: 2006.03866. URL: https://arxiv.org/abs/2006.03866.
- Yang, Zhilin et al. (2019). “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Neural Information Processing Systems*. URL: https://api.semanticscholar.org/CorpusID:195069387.