Универсальная мультифрагментная разметка текстов методом попарной классификации

Н. А. Кукузей*, О. Ю. Седельникова‡, И. А. Цыбанов§, О. Л. Ринк¶, А. И. Майсурадзе‖

Предложен метод двухуровневой матричной классификации для выделения фрагментов текста и объединения их в элементы разметки. Метод допускает пересечение и вложенность фрагментов. Проведён анализ полученных результатов на 4 конфигурациях, включающих 3 различных кодировщика и большую языковую модель. Сравнение моделей производится с помощью критериев качества, вычисляемых на основе сопоставления разметок.

Ключевые слова: двухуровневая разметка текста, попарная классификация.

1. Постановка задачи

Рассматривается задача структурированной разметки текста: выделение фрагментов и объединение их в элементы, причём оба уровня допускают несколько меток. Используется модель данных [1], применимая к извлечению информации и контент-анализу.

Документ состоит из L символов. *Помеченный фрагмент* — пара $f = (I_f, T_f)$, где $I_f = [b, e)$ — непустой полуинтервал, $0 \le b < e \le L$, $b, e \in \mathbb{Z}$, а $T_f \subseteq \mathcal{T}_F$ — набор меток. Множество фрагментов обозначим F.

* *Кукузей Никита Александрович* — Институт ИИ МГУ, e-mail: kukuzeyna@my.msu.ru, ORCID: 0009-0004-0420-9685.

Kukuzei Nikita Alexandrovich — Laboratory Assistant, IAI MSU.

‡ *Седельникова Ольга Юрьевна* — Институт ИИ МГУ, e-mail: sedelnikowaolia@yandex.ru, ORCID: 0009-0007-4923-3701.

Sedelnikova Olga Yuryevna — Laboratory Assistant, IAI MSU.

§ *Цыбанов Илья Антонович* — Стажер-исследователь Института ИИ МГУ, e-mail: tsybanovia@my.msu.ru, ORCID: 0009-0000-1845-464X.

Tsybanov Ilya Antonovich — Research Intern, IAI MSU.

¶ *Ринк Ольга Леонидовна* — н.с. Института ИИ МГУ, e-mail: rinkol@my.msu.ru, ORCID: 0000-0001-5977-6741.

Rink Olga Leonidovna — Researcher, IAI MSU.

‖ *Майсурадзе Арчил Ивериевич* — к. ф.-м. н., в. н. с. Института ИИ МГУ, e-mail: maysuradze@cs.msu.ru, ORCID: 0000-0002-1757-556X.

Maysuradze Archil Iverievich — Cand. of Phys. and Math. Sc., Principal Researcher, IAI MSU.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

Помеченный элемент — пара $e = (S_e, T_e)$, где $\varnothing \neq S_e \subseteq F$ — множество фрагментов, $T_e \subseteq \mathcal{T}_E$ — набор меток элемента. Словари $\mathcal{T}_F, \mathcal{T}_E$ задаются схемой. Выход модели — множества фрагментов F и элементов E.

2. Архитектура

W2NER [2] и T^2-NER [3] решают подзадачу извлечения сущностей; для полной двухуровневой разметки требуется их адаптация. Прямое сравнение не проводилось. В нашей схеме фрагменты и элементы переменного размера имеют независимые наборы меток. Для каждого набора данных обучаются два классификатора поверх замороженного кодировщика, строящего векторы $h_i \in \mathbb{R}^d$ токенов $x_1, \ldots, x_n$. Множество токенных позиций A_f, соответствующих интервалу I_f, определяется по смещениям токенизатора.

$$p_{ij}^{(\ell)} = \sigma(\mathrm{MLP}_\ell([u_i; u_j; u_i \odot u_j; q_{ij}])), \quad \ell \in \{F, E\}.$$

MLP содержит три линейных слоя с размерами выходов d, $d/2$ и числом меток. После первого применяются BatchNorm, GELU и Dropout 0.2, после второго — GELU.

На первом уровне $u_i = h_i$, $q_{ij} = |i - j|$; рассматриваются пары $i < j$, $j - i \leq 64$. На втором u_f — покомпонентный максимум $\{h_i : i \in A_f\}$. Для $b_f = \min A_f$, $r_f = \max A_f + 1$, $m_f = |A_f|$ используются признаки

$$q_{fg} = (|b_f - b_g|, |r_f - r_g|, |m_f - m_g|).$$

Целевая метка $y_{ij,t}^F = 1$, если токены i, j принадлежат одному эталонному фрагменту с меткой t; $y_{fg,t}^E = 1$, если фрагменты f, g входят в один эталонный элемент с меткой t, 0 иначе. Оптимизатор — AdamW с $lr = 10^{-4}$, функция потерь — бинарная кросс-энтропия. Второй уровень обучается на эталонных фрагментах, а применяется к предсказанным.

Пороги каналов максимизируют валидационный F1 на равномерных сетках: $[0.05, 0.95]$ (30 точек) и $[0.05, 0.8]$ (20 точек) для уровней 1 и 2.

Для каждой метки пары $i < j$ с вероятностью выше порога задают рёбра неориентированного графа. Компоненты связности образуют фрагменты или элементы; метки совпадающих компонент объединяются. На втором уровне рассматриваются только пары с $|b_f - b_g| \leq 128$.

3. Результаты

Предложенная архитектура исследована с кодировщиками mamba-1.4b-ru, mDeBERTa-v3-base и spanbert-base-cased. Для сравнения использована

генеративная модель Qwen3.5-397B-A17B-FP8 с примерами в запросе и выводом JSON. Эксперименты охватывают 21 набор данных (NEREL, SciERC, DDI, CoNLL-2012 и др.) и 15 типов задач, включая извлечение именованных сущностей и отношений.

Таблица 1: Качество разметки: точечные оценки, 95%-ные доверительные интервалы. Также дано число параметров модели.

модель	C_1	C_2	C_3	C_4	C_5	размер
Mamba	.110	.430	.783	.306	.778	1,4 млрд
	[.109;.110]	[.428;.432]	[.779;.788]	[.302;.310]	[.774;.783]	
mDeBERTa	.109	.570	.691	.293	.593	279 млн
	[.108;.110]	[.567;.573]	[.687;.695]	[.288;.297]	[.589;.598]	
SpanBERT	.116	.500	.597	.289	.561	110 млн
	[.116;.117]	[.496;.503]	[.593;.601]	[.285;.294]	[.556;.565]	
Qwen	.604	.750	.769	.533	.796	397 млрд
	[.597;.612]	[.741;.757]	[.760;.777]	[.525;.541]	[.787;.803]	

Примечание. (1) Критерии C_1–C_5 взяты из статьи [4]. C_1 характеризует согласованность сопоставления фрагментов; C_2 — согласованность выделенных текстовых интервалов; C_3 — согласованность меток фрагментов; C_4 — согласованность сопоставления элементов разметки; C_5 — согласованность меток элементов разметки. (2) Доверительные интервалы получены процентильным бутстрепом среднего по документам (5000 реплик). Все документы из наборов данных объединяются в один пул с равным весом. Для каждого документа учитывается только одна экспертная разметка.

Качество зависит от кодировщика, однако различия размеров, языков и предобучения не позволяют изолировать эффект архитектуры. Низкие C_1, C_4 указывают на необходимость улучшения выделения фрагментов и восстановления пересекающихся элементов.

Список литературы

[1] A. I. Maysuradze, A. M. Tabachenkov, O. L. Rink, K. V. Vorontsov, “Generalized Text Markup Structures to Organize Knowledge in AI Applications for Humanities and Social Sciences”, *Lobachevskii Journal of Mathematics*, **47**:6 (2026), 2861–2874. DOI: 10.1134/S1995080226619302.

[2] J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, F. Li, “Unified Named Entity Recognition as Word-Word Relation Classification”, *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**:10 (2022), 10965–10973. DOI: 10.1609/aaai.v36i10.21344.

[3] P. Huang, X. Zhao, M. Hu, Z. Tan, W. Xiao, “T^2-NER: A Two-Stage Span-Based Framework for Unified Named Entity Recognition with Templates”, *Transactions of the Association for Computational Linguistics*, **11** (2023), 1265–1282. DOI: 10.1162/tacl_a_00602.

[4] A. Levikin, I. Khabutdinov, A. Grabovoy, K. Vorontsov, “The methodology of multi-criteria evaluation of text markup models based on inconsistent expert markup”, *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2025). Issue 23, Supplementary volume*, 2025, 1066–1080. DOI: 10.28995/2075-7182-2025-23-1066-1080.

Статья поступила 23 октября 2026 г.

Universal Multi-Span Text Annotation through Pairwise Classification

N. A. Kukuzei, O. Y. Sedelnikova, I. A. Tsybanov, O. L. Rink, A. I. Maysuradze

We propose a two-level grid-tagging method for identifying text spans and grouping them into annotation elements. The method supports overlapping and nested spans. We provide an analysis of results obtained with four configurations comprising three different text representation models and a large language model. Performance is evaluated using metrics computed by annotation matching.

Keywords: two-level text annotation, pairwise classification.

References

[1] A. I. Maysuradze, A. M. Tabachenkov, O. L. Rink, K. V. Vorontsov, “Generalized Text Markup Structures to Organize Knowledge in AI Applications for Humanities and Social Sciences”, *Lobachevskii Journal of Mathematics*, **47**:6 (2026), 2861–2874. DOI: 10.1134/S1995080226619302.

[2] J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, F. Li, “Unified Named Entity Recognition as Word-Word Relation Classification”, *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 2022, 10965–10973. DOI: 10.1609/aaai.v36i10.21344.

[3] P. Huang, X. Zhao, M. Hu, Z. Tan, W. Xiao, “T^2-NER: A Two-Stage Span-Based Framework for Unified Named Entity Recognition with Templates”, *Transactions of the Association for Computational Linguistics*, **11** (2023), 1265–1282. DOI: 10.1162/tacl_a_00602.

[4] A. Levikin, I. Khabutdinov, A. Grabovoy, K. Vorontsov, “The methodology of multi-criteria evaluation of text markup models based on inconsistent expert markup”, *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2025). Issue 23, Supplementary volume*, 2025, 1066–1080. DOI: 10.28995/2075-7182-2025-23-1066-1080.

Received on October 23, 2026