Эффективный аудит текстовых разметок на основе индивидуальных оценок сложности

А. М. Табаченков*, А. И. Майсурадзе‡

Аудит текстовой разметки, включающий обнаружение и исправление ошибочных границ и меток фрагментов, играет ключевую роль в социально-гуманитарных исследованиях, однако его автоматизация требует существенных затрат. Для снижения затрат предлагается метод на основе оценок сложности с эффективной векторизацией и однократным кодированием документов. Сравнение с LLM на трех источниках исправлений показывает сопоставимое качество в ряде задач аудита при значительно меньшем времени работы.

Ключевые слова: аудит разметки; текстовые фрагменты; оценки сложности; коррекция разметки; LLM-проверка.

1. Мотивация и цель исследования

Надежность данных и моделей зависит от качества разметки. Однако неоднозначность и редкие метки [2] затрудняют разметку, а аудит [1] и его автоматизация затратны. Так, большие языковые модели (LLM) [4] требуют генерации ответов, итеративное самообучение [3] — повторного обучения. Диффузионный метод [5] уточняет предсказания NER, а не человеческую разметку. Поэтому наша цель — разработать легковесный System 1-подход, сопоставимый с LLM по качеству.

2. Постановка задачи и типы ошибок

Разметка — набор символьных фрагментов $[b, e)$ с метками Y. Детекция выявляет ошибки [1]; коррекция исправляет либо удаляет фрагменты.

* *Табаченков Андрей Михайлович* — Лаборант, Институт ИИ МГУ, e-mail: tabachenkovam@my.msu.ru, ORCID: 0009-0005-4980-1065.

Tabachenkov Andrey Mikhailovich — Laboratory Assistant, 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

3. Предлагаемый метод

Мы предлагаем использовать схему [6]: кодировщик обучается классификации фрагментов по исходной разметке без эталонных исправлений. Документ кодируется однократно; вектор фрагмента объединяет представления граничных токенов, среднее и максимум. Согласно [6], уверенность, отступ, расстояние Махаланобиса и дисперсия градиентов (VoG) отражают сложность фрагментов. Для детекции пороги оценок и их процентильных комбинаций максимизируют TPR−FPR на синтетических искажениях обучающей выборки. При коррекции выбираются метка, затем локальные границы. Отрицательная O означает удаление фрагмента, а в многометочном случае — проверяемой метки. В многометочном случае каждая пара фрагмент — метка проверяется отдельно. Замена принимается по порогу совместной оценки подозрительности и уверенности с учетом процентильного выигрыша кандидата.

Затраты имеют вид $T_d \approx T_{\text{enc}}(d) + K_d T_{\text{score}}$, где K_d — число кандидатов, T_{enc} — время кодирования, T_{score} — векторизации, классификации и оценки кандидата. VoG требует пяти кодирований и градиентов по векторам фрагментов; переобучение не требуется.

4. Схема экспериментов

Существующие бенчмарки [1] сосредоточены преимущественно на детекции. Рассматриваются три источника исправлений: синтетические искажения, реальные исправления и консенсус разметчиков. Из 5934 фрагментов проверочной части английского CoNLL-2003 половина сохраняется, остальные равновероятно получают сдвиг границ до двух символов, другую положительную метку либо оба искажения; для удаления добавляются 297 ложных фрагментов. В CleanCoNLL [7] венгерский алгоритм сопоставляет исходные и исправленные фрагменты по пересечению; изолированная правка границ сохраняет метку, несопоставленные исходные фрагменты считаются удаленными. В Human Values [2] аналогично сопоставляются фрагменты остальных экспертов по метке: эталон задается строгим большинством, границы — медоидом; неподдержанные фрагменты удаляются, недостающие метки добавляются. Разбиение CoNLL соответствует [7], Human Values — как при обучении кодировщика. Эталонные исправления на контрольной выборке служат лишь для оценки качества детекции и коррекции.

Мы сравниваем DeBERTa [6] и Qwen3.5-4B/9B (условный System 2). Вход LLM [4] — задача, метки и фрагмент с контекстом; выход — ошибки

и исправления в JSON. Генерация Qwen жадная, без рассуждений. Эксперименты проведены на NVIDIA Tesla T4 с 16 ГБ видеопамяти. Детекция оценивается по ROC-AUC, AP, precision, recall и F1 [1]; коррекция — по совпадению границ и метки, IoU и отклонению границ.

5. Итоги

На 6231 синтетическом примере CoNLL кодирование и оценка кандидатов DeBERTa занимают около 35 с, отдельно VoG — 82 с, генерация Qwen-4B/9B — 1 ч 9 мин / 3 ч 13 мин (без обучения, настройки и загрузки моделей). При этом на ошибках границ CleanCoNLL усреднение процентилей пяти оценок сложности дает 64,6% точных исправлений против 35,4% у Qwen-9B; на синтетических ошибках меток — 97,1% для DeBERTa и 84,4% для Qwen-9B; на Human Values VoG обнаруживает ошибки границ с ROC-AUC 0,694. Однако на реальных ошибках меток CleanCoNLL Qwen-9B дает 54,8% против 8,8% у DeBERTa. Приведенные результаты коррекции получены при известном действии и не характеризуют качество полного автоматического аудита. Таким образом, System 1 позволяет снизить затраты на аудит без потери качества в ряде задач, однако преимущество того или иного подхода зависит от типа ошибки.

Список литературы

[1] J.-C. Klie, B. Webber, I. Gurevych, “Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future”, *Computational Linguistics*, **49**:1 (2023), 157–198. DOI: 10.1162/coli_a_00464.

[2] O. Rink et al., “Automated Detection of Human Values in Texts: ML Challenges and Performance Benchmarks”, *Social Computing and Social Media*, 2025, 304–321. DOI: 10.1007/978-3-031-93536-7_22.

[3] T. Ehsan, T. Solorio, “A Scalable Framework for Automated NER Annotation Correction in Low-Resource Languages”, *Findings of EACL*, 2026, 4138–4151. DOI: 10.18653/v1/2026.findings-eacl.215.

[4] H. Kim et al., “MEGAnno+: A Human-LLM Collaborative Annotation System”, *EACL: System Demonstrations*, 2024, 168–176. DOI: 10.18653/v1/2024.eacl-demo.18.

[5] L. H. Trieu, H. M. Phi, M. Miwa, “Reviving Iterative Refinement in Diffusion-based NER with an Initializer-Restorer Approach”, *ACL*, 2026, 877–882. DOI: 10.18653/v1/2026.acl-short.72.

[6] A. I. Maysuradze, A. M. Tabachenkov, “Adaptation and Interpretation of Various Instance Hardness Assessments for Text Spans”, *Lobachevskii Journal of Mathematics*, **47**:6 (2026), 2845–2860. DOI: 10.1134/S1995080226618540.

[7] S. Rücker, A. Akbik, “CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset”, *EMNLP*, 2023, 8628–8645. DOI: 10.18653/v1/2023.emnlp-main.533 .

Статья поступила 24 октября 2026 г.

Efficient Audit of Text Annotations Using Instance Hardness Assessments

A. M. Tabachenkov, A. I. Maysuradze

Detecting and correcting span boundary and label errors is essential in social sciences and humanities but costly to automate. To reduce costs, we propose a method using instance hardness assessments, efficient vectorization, and one-time document encoding. Comparison with LLMs on three correction sources shows comparable quality on several auditing tasks with substantially shorter runtimes.

Keywords: annotation auditing; text spans; instance hardness; annotation correction; LLM-as-a-judge.

References

[1] J.-C. Klie, B. Webber, I. Gurevych, "Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future", *Computational Linguistics*, **49**:1 (2023), 157–198. DOI: 10.1162/coli_a_00464.

[2] O. Rink et al., "Automated Detection of Human Values in Texts: ML Challenges and Performance Benchmarks", *Social Computing and Social Media*, 2025, 304–321. DOI: 10.1007/978-3-031-93536-7_22.

[3] T. Ehsan, T. Solorio, "A Scalable Framework for Automated NER Annotation Correction in Low-Resource Languages", *Findings of EACL*, 2026, 4138–4151. DOI: 10.18653/v1/2026.findings-eacl.215.

[4] H. Kim et al., "MEGAnno+: A Human-LLM Collaborative Annotation System", *EACL: System Demonstrations*, 2024, 168–176. DOI: 10.18653/v1/2024.eacl-demo.18.

[5] L. H. Trieu, H. M. Phi, M. Miwa, "Reviving Iterative Refinement in Diffusion-based NER with an Initializer-Restorer Approach", *ACL*, 2026, 877–882. DOI: 10.18653/v1/2026.acl-short.72.

[6] A. I. Maysuradze, A. M. Tabachenkov, "Adaptation and Interpretation of Various Instance Hardness Assessments for Text Spans", *Lobachevskii Journal of Mathematics*, **47**:6 (2026), 2845–2860. DOI: 10.1134/S1995080226618540.

[7] S. Rücker, A. Akbik, "CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset", *EMNLP*, 2023, 8628–8645. DOI: 10.18653/v1/2023.emnlp-main.533.

Received on October 24, 2026