



ИИ Институт
мгу искусственного
интеллекта

Методология многокритериальной оценки текстовой разметки моделей на основе несогласованной экспертной разметки

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН

k.vorontsov@iai.msu.ru

Конкурс ПРО//ЧТЕНИЕ

(<http://ai.upgreat.one>)



Задача: автоматическая разметка смысловых ошибок в сочинениях ЕГЭ по русскому яз., литературе, истории, обществознанию, английскому яз.

Период: декабрь 2019 — июнь 2022, три цикла испытаний.

Призовой фонд: ₹100M русский язык + ₹100M английский язык

Типов ошибок: 152

(р:70 л:16 о:23 и:20 а:23)

Подтипов ошибок: 236

(р:112 л:19 о:29 и:26 а:50)

Помимо выделения ошибок, надо давать их объяснения.

ФАКТИЧЕСКАЯ ОШИБКА

автор высказывания А.Франц

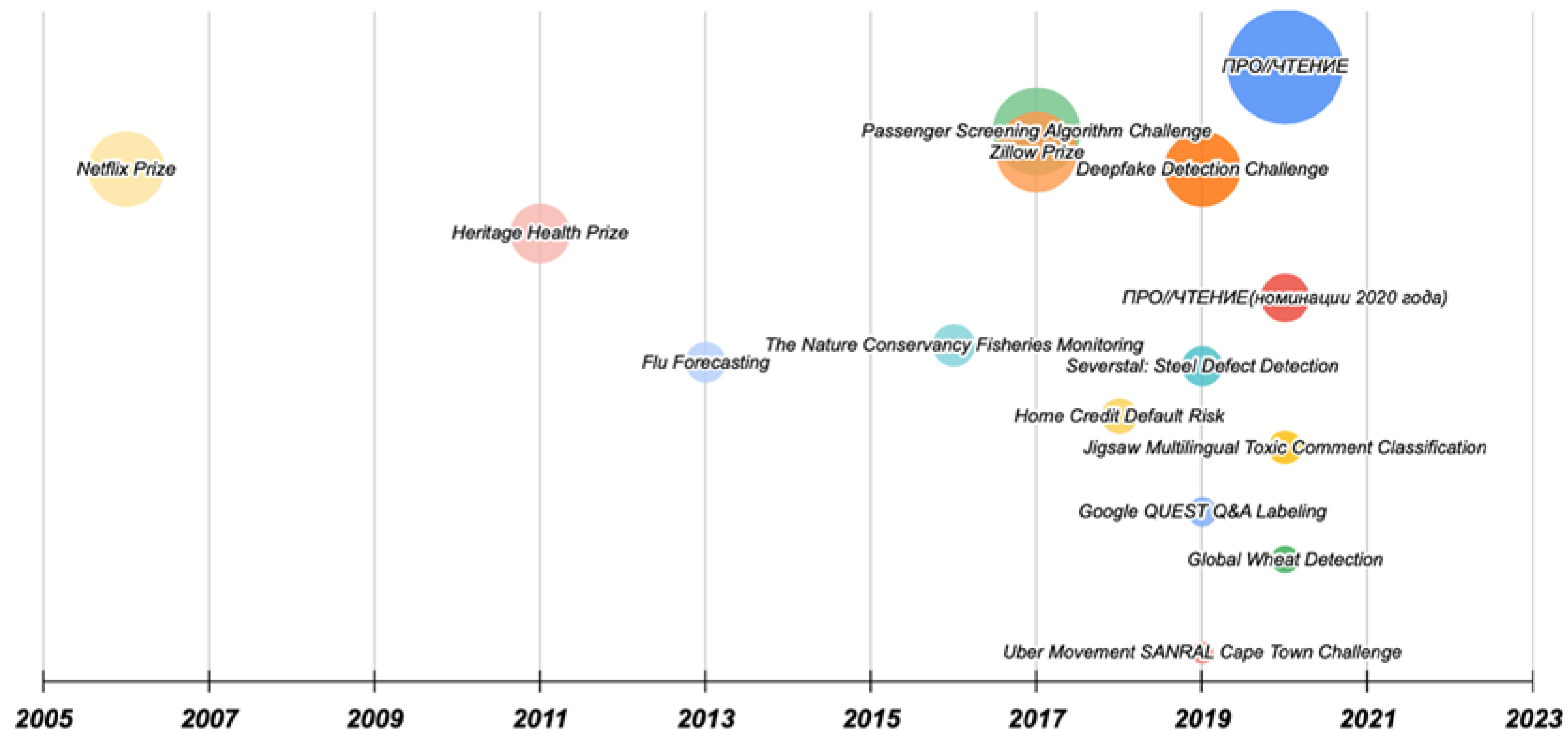
В своем высказывании «Если человек зависит от природы, то и она от него зависит» Д. Мережковский говорит о необходимости защиты природы.

ЛОГИЧЕСКАЯ ОШИБКА

тезис не обоснован

Конкурс ПРО//ЧТЕНИЕ

(<http://ai.upgreat.one>)



Конкурс ПРО//ЧТЕНИЕ

(<http://ai.upgreat.one>)



Сравнение двух разметок:

Алгоритмическая разметка

Нередко люди совершают плохие поступки, забывая о том, что, даже скрыв свой поступок от других, человек не скроется от своей совести. Что же такое безнравственный поступок? Безнравственный поступок - это поступок, не соответствующий моральным нормам.

Можно ли оправдать безнравственный поступок? Именно эту проблему В. Ф. Тендряков поднимает в **своем** тексте. Докажем сказанное примерами из представленного отрывка.

В тексте В. Ф. Тендряков говорит о том, что человек во благо себе может легко совершить низкий поступок, не испытав при этом чувство стыда. Человек сможет оправдать свой поступок перед самим собой, объяснив причину. В пример автор приводит поведение героя, который часто в жизни совершал безнравственные поступки. Он врал, дрался и крал. Мы видим, что до войны герой привык совершать плохие поступки. Он всегда оправдывался, потому что не хотел нести ответственность за свои действия, а значит не испытывал мучения совести. Мы знаем, что муки совести – это первое и самое сильное наказание, которое получает человек, совершивший плохой поступок. Но наш герой не получал никакого наказания и поэтому продолжал совершать безнравственные поступки. Проанализировав поведение главного героя, я убедилась в том, что человек обязан нести ответственность за свои поступки всегда, и поэтому я утверждаю, что нельзя оправдывать даже мелкие безнравственные поступки.

СВЯЗЬ РПОВТОР

РПОВТОР РЛИШН ПРОБЛЕМА

РПОВТОР РПОВТОР РПОВТОР

РЛИШН

РПОВТОР

РПОВТОР

РПОВТОР

РПОВТОР ГОДНОР ГОДНОР

РВИДОВР РПОВТОР

РПОВТОР РПОВТОР

РПОВТОР РПОВТОР

РПОВТОР РВИДОВР РПОВТОР

РПОВТОР

РПОВТОР

Экспертная разметка 2

Нередко люди совершают плохие поступки, забывая о том, что, даже скрыв свой поступок от других, человек не скроется от своей совести. Что же такое безнравственный поступок? Безнравственный поступок - это поступок, не соответствующий моральным нормам.

Можно ли оправдать безнравственный поступок? Именно эту проблему В. Ф. Тендряков поднимает в своем тексте. Докажем сказанное примерами из представленного отрывка.

В тексте В. Ф. Тендряков говорит о том, что человек во благо себе может легко совершить низкий поступок, не испытав при этом чувство стыда. Человек сможет оправдать свой поступок перед самим собой, объяснив причину. В пример автор приводит поведение героя, который часто в жизни совершал безнравственные поступки. Он врал, дрался и крал. Мы видим, что до войны герой привык совершать плохие поступки. Он всегда оправдывался, потому что не хотел нести ответственность за свои действия, а значит не испытывал мучения совести. Мы знаем, что муки совести – это первое и самое сильное наказание, которое получает человек, совершивший плохой поступок. Но наш герой не получал никакого наказания и поэтому продолжал совершать безнравственные поступки. Проанализировав поведение главного героя, я убедилась в том, что человек обязан нести ответственность за свои поступки всегда, и поэтому я утверждаю, что нельзя оправдывать даже мелкие безнравственные поступки.

P100STOP T1
 P100STOP T1
 P100STOP T2 P100STOP T1
 P100STOP T2 P100STOP T2
 P100STOP T3
 P100STOP T4 P100STOP T1 P100STOP T1
 P100STOP T1
 P100STOP T4
 P100STOP T1
 P100STOP T4 P100STOP T1
 P100STOP T4 P100STOP T1
 P100STOP T4 P100STOP T1
 P100STOP T1
 P100STOP T1

Оценивание алгоритмической разметки



- В основе методики — парное сравнение разметок текста:
«алгоритм $\leftrightarrow$ эксперт», «эксперт-1 $\leftrightarrow$ эксперт-2»
на основе оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок $C_n(A, B)$, $n = 1, \dots, N$
- Вводится их средневзвешенная согласованность $Con(A, B)$
- СТАР (Средняя Точность Алгоритмической Разметки)
— средняя по выборке $Con(A, E)$ разметок алгоритма A и эксперта E
- СТЭР (Средняя Точность Экспертной Разметки)
— средняя по выборке $Con(E_1, E_2)$ разметок двух экспертов, E_1 и E_2
- ОТАР = СТАР / СТЭР, если больше 100%, то модель лучше экспертов

Критерии согласованности разметок

Оптимальное сопоставление элементов разметок A и B



Критерии (числовые величины от 0 до 1; чем выше, тем лучше):

C_1 = доля фрагментов, для которых найдено сопоставление

C_2 = точность наложения сопоставленных фрагментов

C_3 = точность совпадения тегов сопоставленных фрагментов

C_4 = точность совпадения связей сопоставленных фрагментов

C_5 = точность совпадения затекстов сопоставленных фрагментов

Контент-анализ: обобщение и автоматизация

Обобщённый контент-анализ — четыре базовые операции с текстом:

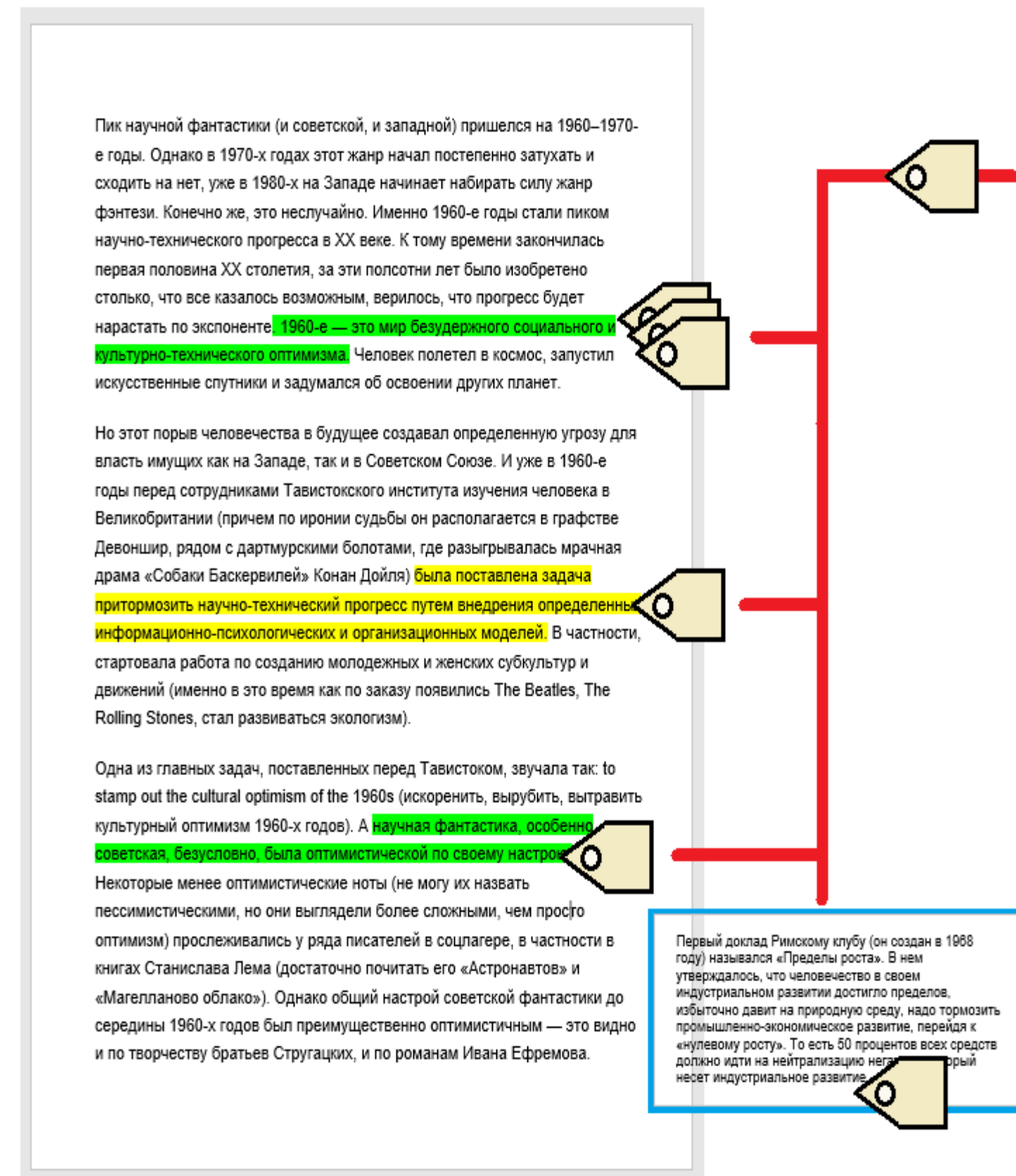
- 1) выделить фрагмент
- 2) классифицировать (тегировать) фрагмент по рубрикатору
- 3) *связать несколько фрагментов*
- 4) *дать комментарий (затекст) к фрагменту или связи*

Цель — автоматизировать контент-анализ больших текстовых массивов по небольшим размеченным корпусам, в любой предметной области

Три задачи построения обучаемой модели разметки:

- 1) разработка рубрикатора и инструкций разметчика
- 2) дообучение большой языковой модели по разметке
- 3) оценивание качества, сравнение и выбор моделей

Разметка текста: обобщённый контент-анализ

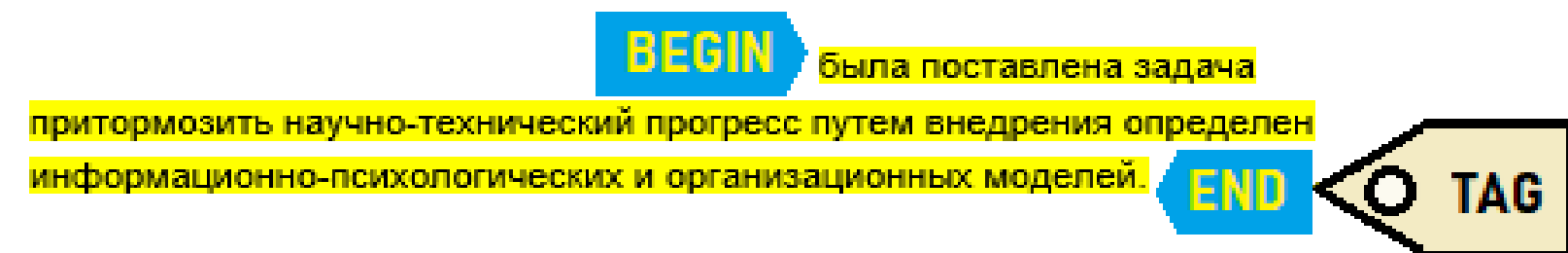


Разметка состоит из элементов

Элемент разметки — несколько взаимосвязанных фрагментов, затекстов и тегов

Теги (классы) выбираются из рубрикатора

Фрагмент задаётся началом и концом, может иметь один или несколько тегов:



Затекст — комментарий, объяснение, дополнительная информация и т.п., может иметь один или несколько тегов

Инструмент разметки

<https://markup.mlsa-iai.ru>

←

Задания

Теги

Документы

Постобработка

Пользователь

🔔 ? ⓘ ↗

➤ Список тегов

Документ №1: "23 февраля в войсковой части 5526..."

23 февраля в войсковой части 5526 прошло чествование воспитанников военно-патриотического клуба «Крепость». Заместитель командира части по идеологической работе майор Олег Ляшук отметил, что воспитание подрастающего поколения в патриотическом ключе, в стремлении к здоровому образу жизни, уважению к традициям, культурным ценностям и исторической памяти государства является главным профилактическим фактором безнравственности и аморальности. «Вместе мы вносим огромный вклад в будущее страны и нравственное здоровье нашего общества. Служба Родине во все времена была почетна. А служить можно по-разному, и не обязательно с оружием в руках. Служить можно в любом возрасте, служить можно и парню, и девушке. Служить можно и словом, и делом!» – подытожил Олег Ляшук.

Добавить фрагмент

Разметка от 09.02.2025: "23 февраля в войсковой части 5526..."

Выбрано объектов: 0

Выбрать все

Убрать все

▼ ☐ Элемент №2

Здоровье × Тональность: Положительная ×

☐ "в стремлении к здоровому образу жизни" ×

▼ ☐ Элемент №2

Уважение традиций × Тональность: Положительная ×

☐ "уважению к традициям" ×

▼ ☐ Элемент №3

Справедливость × Тональность: Положительная ×

☐ "служить можно по-разному, и не обязательно с оружием в руках. Служить можно в любом возрасте, служить можно и парню, и девушке. Служить можно и словом, и делом!" ×

Сохранить и выйти

Список публикаций

1. *Rink O.L., Lobachev V.A., Vorontsov K.V.* Detecting human values and sentiments in large text collections with a context-dependent information markup: a methodology and math. HCII 2024.
2. *Vorontsov K.V., Gladchenko I.A., Lobachev V.A., Mamontova A.V., Rink O.L., Shabelskaya N.K.* Methodology for detecting human values in large text collections // Bulletin of St. Petersburg University. International relations. 2024.
3. *Rink O. L., Vorontsov K. V., Shabelskaya N. K.* Uncovering positivism, negativism, and conflict in large text collections. Material values and the code of «noble maidens» // EDN ADEUQP, April 18–20, 2024. – P. 137-139.
4. *Maysuradze A.I., Rink O.L., Fedorov A.M., Tabachenkov A.M., Vorontsov K.V.* Does annotating multi-spans improve classification in large text collections? 2024ICSAI (China & IEEE). Lecture Notes in Computer Science series, 2024.
5. *Vorontsov K.V., Gladchenko I.A., Lobachev V.A., Mamontova A.V., Rink O.L., Shabelskaya N.K.* Developing an Open Interdisciplinary Classifier of Human Values by means of Annotating Multi-fragments // Bulletin of St. Petersburg University. International relations. 2025.
6. *Rink O.L., Maysuradze A.I., Fedorov A.M., Ischenko R.V., Korchagina A.V., Tabachenkov A.M., Tsybanov I.A., Vorontsov K.V.* Automated detection of human values in texts: ML challenges and performance benchmarks. 2025.

Методология многокритериальной оценки текстовой разметки моделей на основе несогласованной экспертной разметки

Грабовой Андрей Валериевич, к.ф.-м.н.
руководитель отдела исследований в компании Антиплагиат

г. Москва, 2025 год

Экспертные модели проверки сочинений 2022

Грамматика и пунктуация

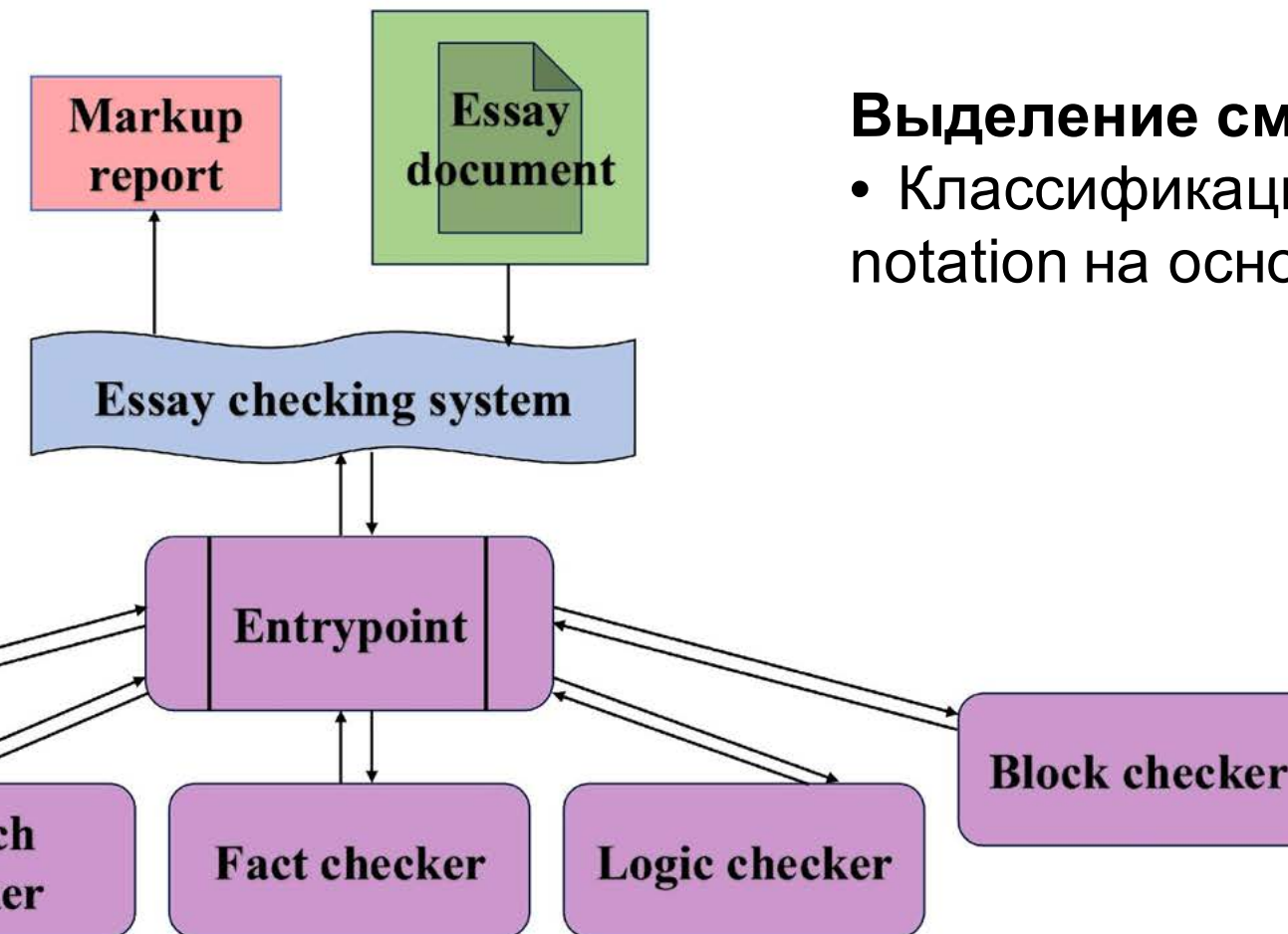
- Адаптация GECToR в RuGECToR [1]
- Набор регулярных выражений [2]

Проверка речевых норм

- Регулярные выражения и ручные правила
- BERT-CRF: классифицируют сложные лексические нарушения

Проверка фактических ошибок

- Алгоритм FEVER для каждого “высказывания”:
 - выделение фактов – BERT класс. предложений
 - поиск кандидатов документов – bm25
 - поиск релевантных фрагментов – LaBSE
 - верификация фрагментов – BERT класс



Выделение смысловых блоков

- Классификация токенов на смысловые блоки BI-notation на основе BERT-CRF модели

Проверка логических ошибок

- Классификация абзацев на связность – NLU BERT
- Проверка логической последовательности предложений – QA-BERT

Генерация пояснений и комментариев

- Каждая ошибка – тег исправления, для которого ставится в соответствие пояснение и комментарий

[1] I. A. Khabutdinov, A. V. Grabovoy, et al. Rugector: Rule-based neural network model for russian language grammatical error correction. *Programming and Computer Software*, 50(4):315–321, 2024.

[2] LanguageTool: <https://github.com/languagetool-org/languagetool>

Развитие моделей проверки сочинений 2025

Грамматика и пунктуация

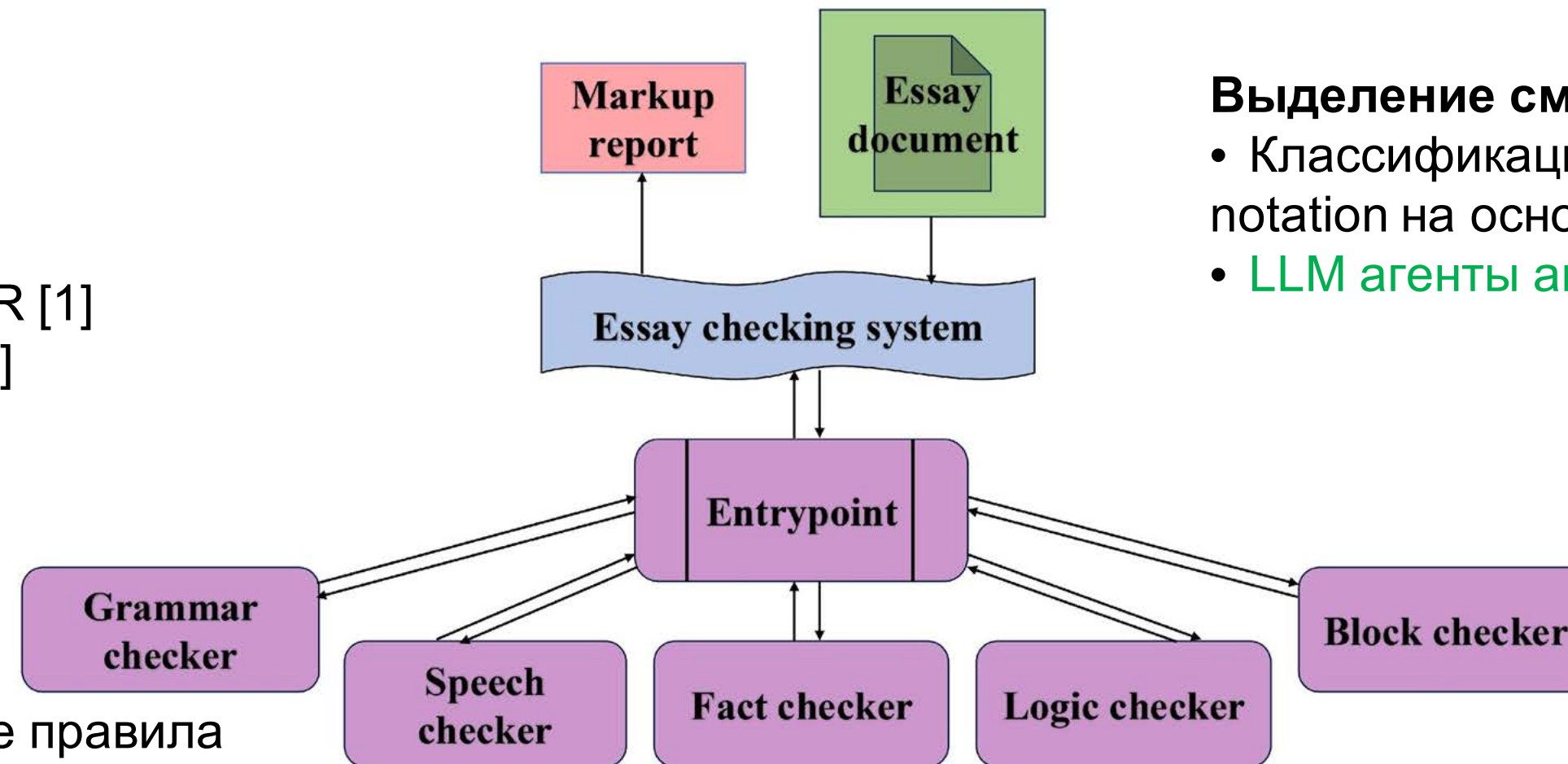
- Адаптация GECToR в RuGECToR [1]
- Набор регулярных выражений [2]
- LLM агенты анализа текста

Проверка речевых норм

- Регулярные выражения и ручные правила
- BERT-CRF: классифицируют сложные лексические нарушения

Проверка фактических ошибок

- Алгоритм FEVER для каждого “высказывания”:
 - выделение фактов – BERT класс. предложений
 - поиск кандидатов документов – bm25
 - поиск релевантных фрагментов – LaBSE
 - LLM агенты анализа текста [3]



Выделение смысловых блоков

- Классификация токенов на смысловые блоки BI-notation на основе BERT-CRF модели
- LLM агенты анализа текста

Проверка логических ошибок

- Классификация абзацев на связность – NLU BERT
- Проверка логической последовательности предложений – QA-BERT

- LLM агенты анализа текста

Генерация пояснений и комментариев

- Каждая ошибка – тег исправления, для которого ставиться в соответствие пояснение и комментарий
- LLM для генерации пояснений на основе тегов

[1] I. A. Khabutdinov, A. V. Grabovoy, et al. Rugector: Rule-based neural network model for russian language grammatical error correction. *Programming and Computer Software*, 50(4):315–321, 2024.

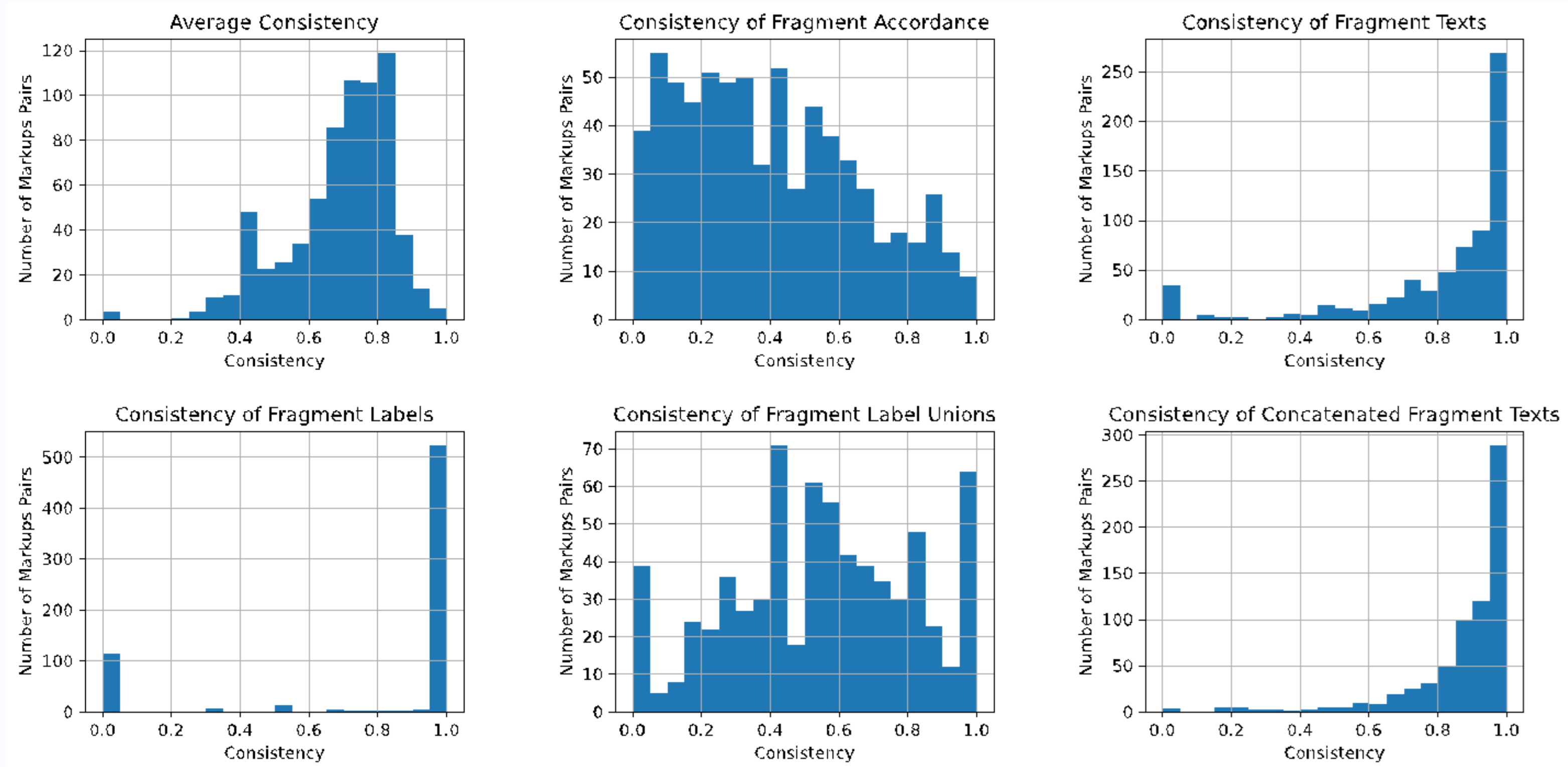
[2] LanguageTool: <https://github.com/languagetool-org/languagetool>

[3] A. Voznyuk, G. Gritsay, A. Grabovoy. Advacheck at SemEval-2025 Task 3: Combining NER and RAG to Spot Hallucinations in LLM Answers. *Proceedings of the 19th IWSE (SemEval-2025)*.

Использование методологии разметки текстов

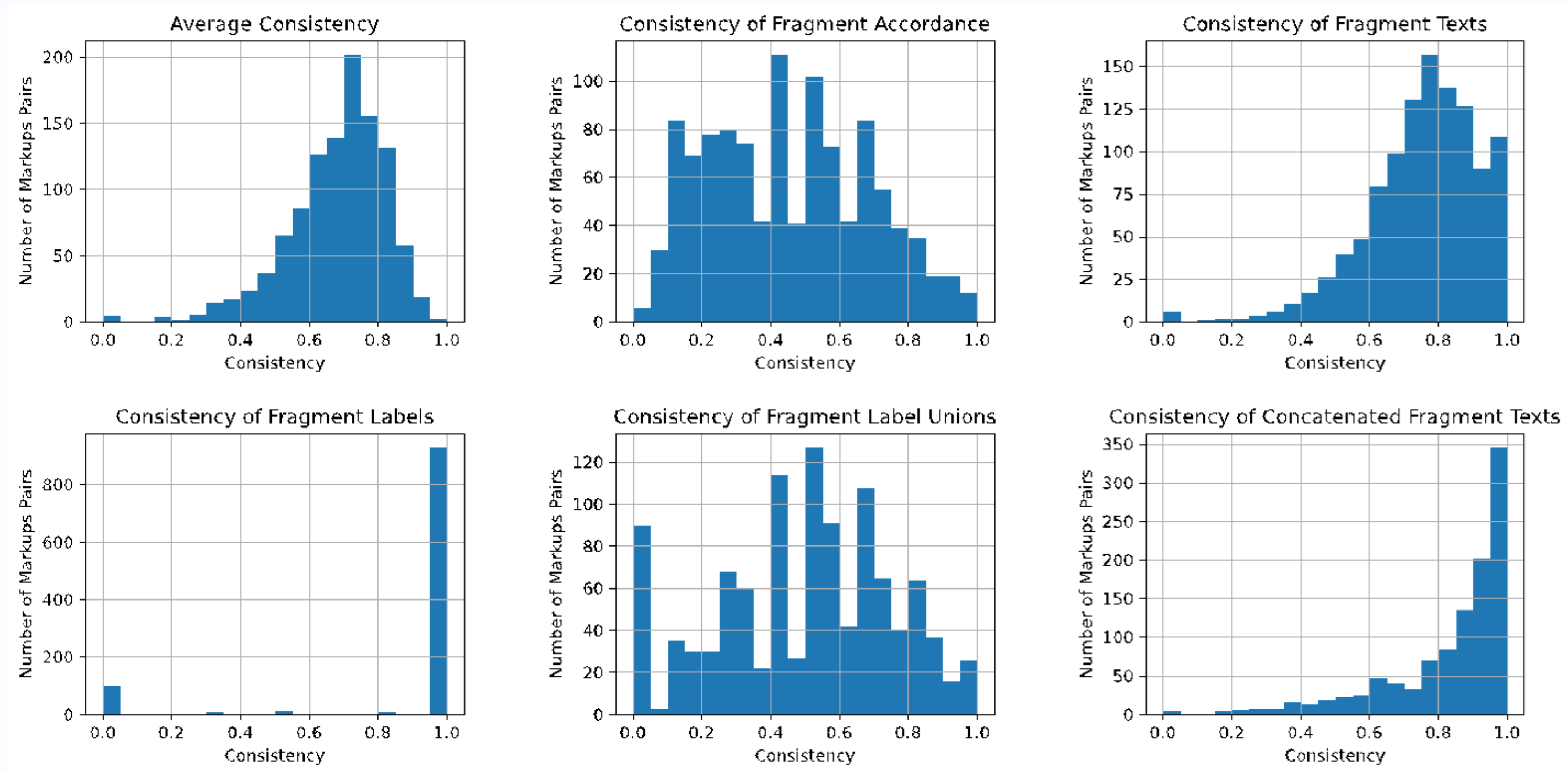
- Формализация метода оценки сменила задачу проверки сочинения на задачу разметки текста:
 - a. решение задачи комплексной оценки сочинения ЕГЭ
 - b. декомпозиция задачи и сбор результата для комплексной оценки
- LLM с заданными правилами разметки текста:
 - a. задает *контракт* взаимодействия с LLM (формат входа-выхода)
 - b. уменьшает вероятность галлюцинаций, задав *рамки* ответа модели
 - c. является частью входных данных, которые определяет работу модели

Критерии согласованности для пар экспертных разметок



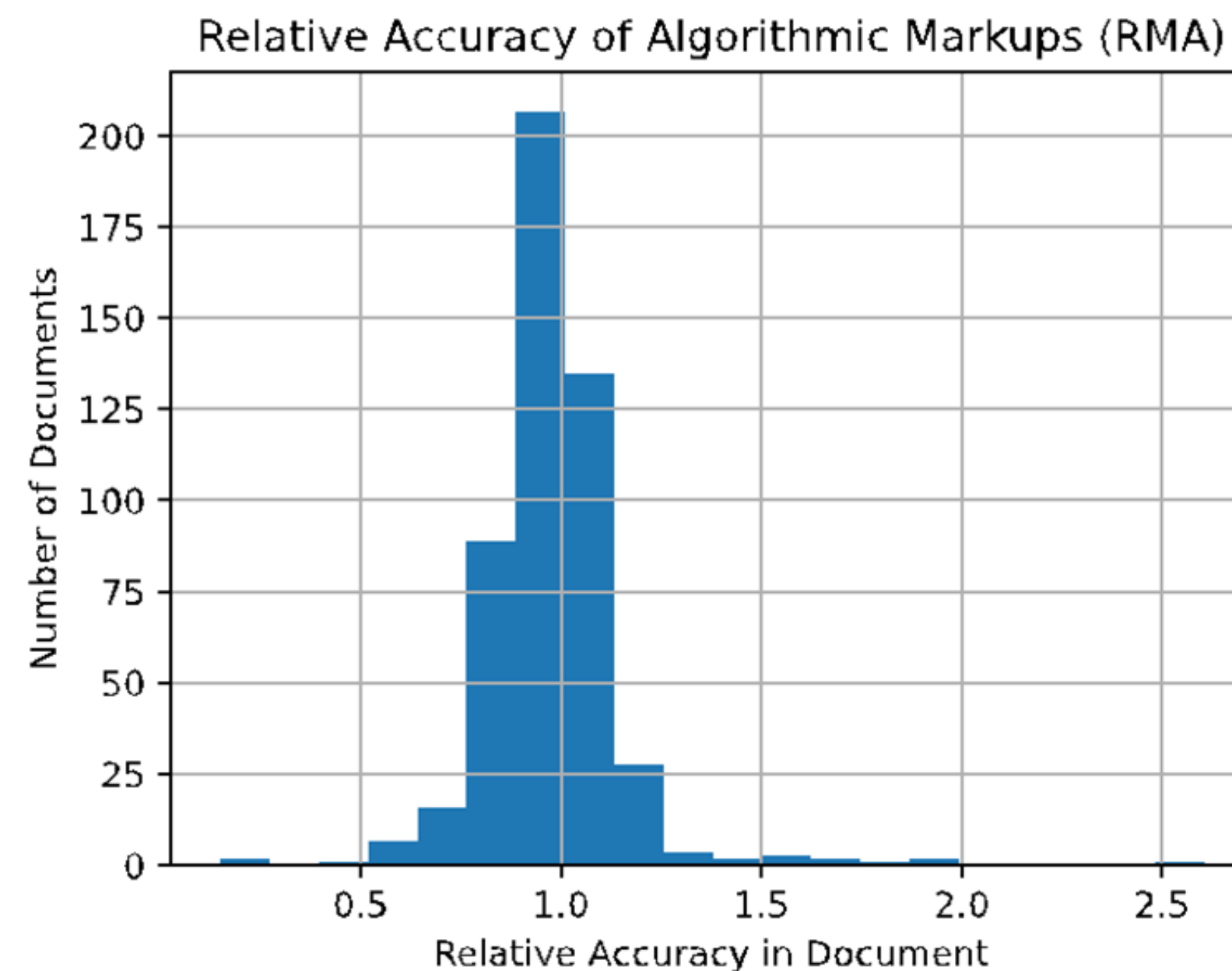
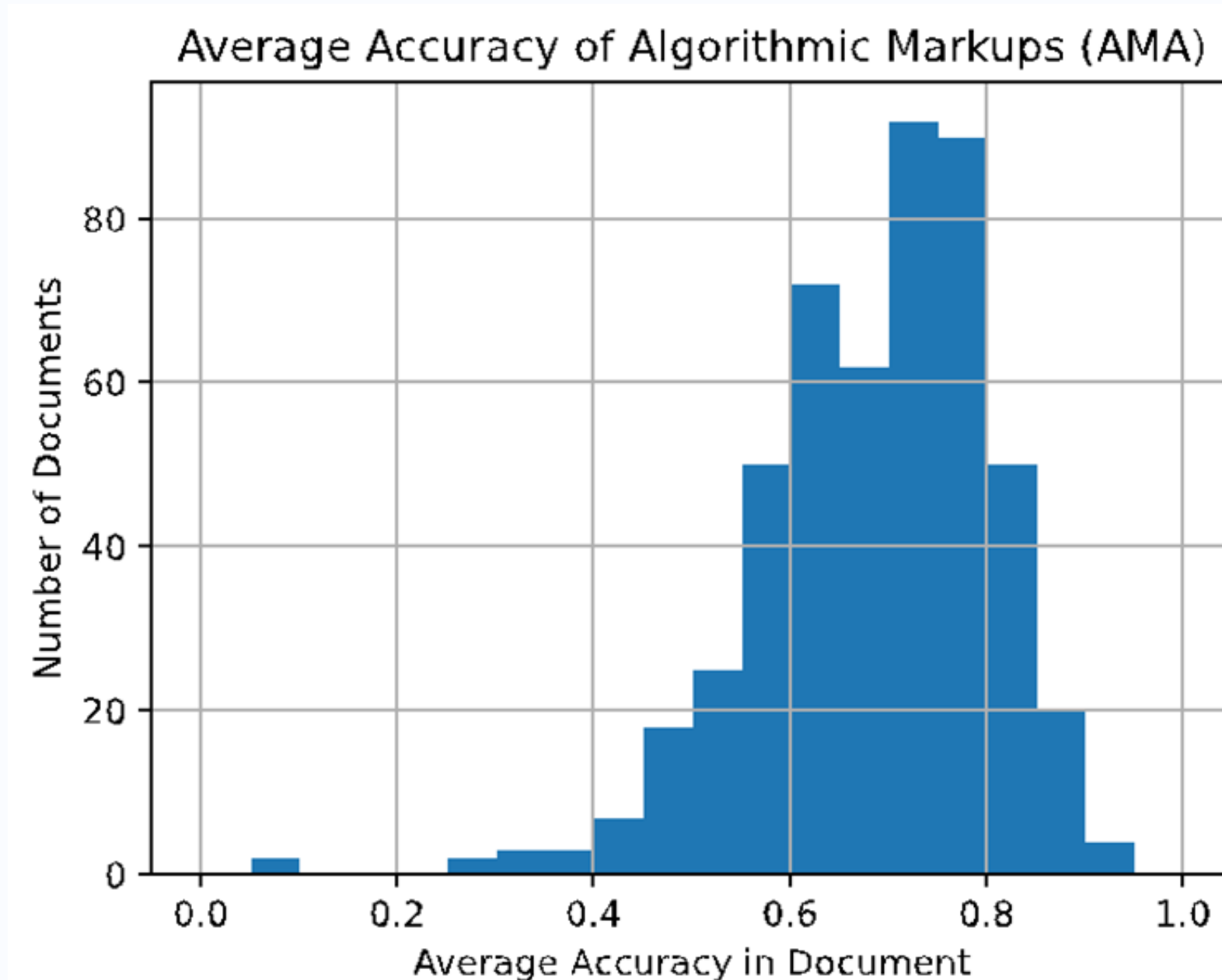
1. Широкое распределение значений указывает на расхождение оценок экспертов.
2. Субъективность задачи разметки эссе и необходимость использования специальных метрик согласованности.
3. При отсутствии "истинной" разметки важно учитывать не абсолютную точность, а степень согласия между экспертами.

Критерии согласованности пар алгоритм-эксперт



1. Алгоритм показывает схожее с человеком качество разметки.
2. Алгоритм демонстрирует более значительный разброс.
3. Указывает на то, что хотя алгоритм эффективен, есть случаи, когда он либо извлекает неправильные фрагменты, либо не может последовательно идентифицировать некоторые ошибки.

Показатели качества модели в рамках конкурса ПРО//ЧТЕНИЕ



1. RMA: в среднем алгоритм имеет такую же точность, как и эксперты между собой, но встречаются крайние случаи — как превосходства алгоритма, так и его слабости.
2. AMA: высокие значения точности, однако небольшое смещение к низким значениям указывает на отдельные случаи, где алгоритм работает с неоднозначностью.
3. Эти результаты подтверждают, что алгоритм в целом сравним с экспертами, но есть возможности для улучшения.