Научный семинар «Проблемы управления знаниями» — ИПУ РАН, 15 октября 2025

От регламента технологического конкурса ПРО//ЧТЕНИЕ к машинному обучению моделей контент-анализа по несогласованным экспертным разметкам

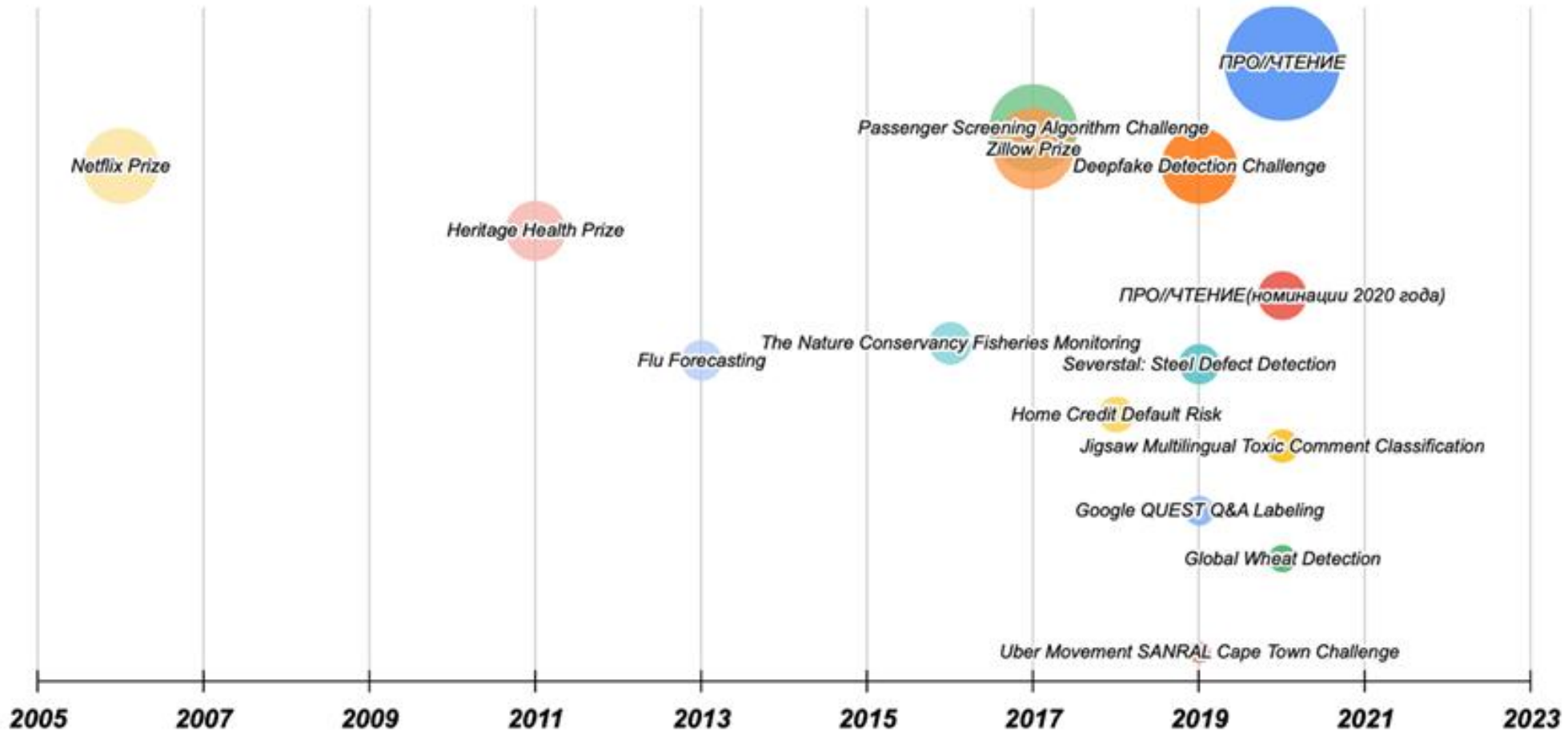
Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,

зав. лабораторией *машинного обучения и семантического анализа*

Института искусственного интеллекта МГУ им. М.В. Ломоносова

Конкурсы анализа данных — драйверы цифрового развития отраслей



Конкурс ПРО//ЧТЕНИЕ
(http://ai.upgreat.one)

Up Great technology contest

READ // ABLE comprehension technology

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

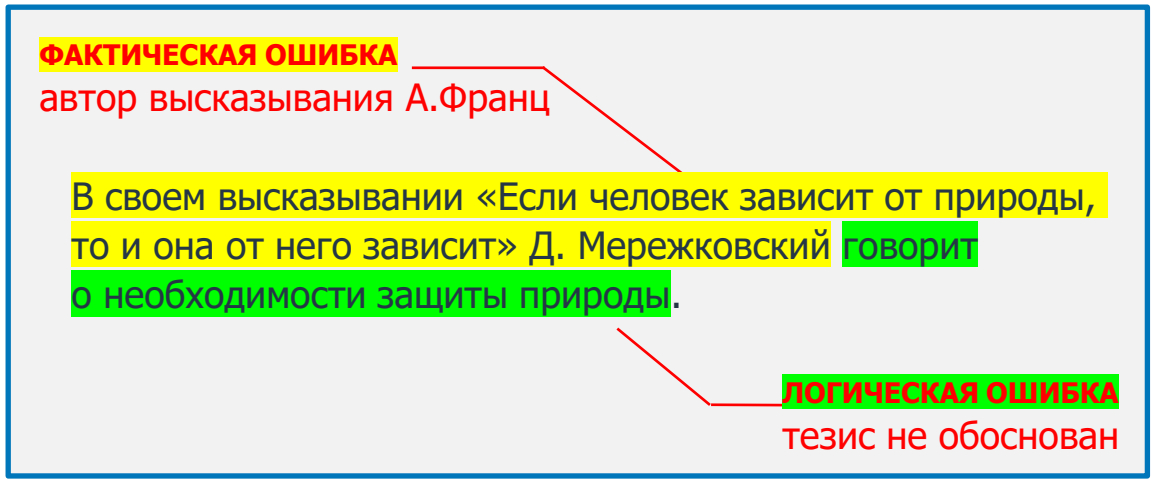
Период: декабрь 2019 — июнь 2022, три цикла испытаний.

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.



Конкурс ПРО//ЧТЕНИЕ
(http://ai.upgreat.one)

Up Great technology contest

READ // ABLE
comprehension technology

Сравнение двух разметок:

ЯЗЫК РАЗМЕТКИ ДАННЫХ

(* Коды \ Фрагмент \ Комментарий :: Пояснение >> Исправление # Тег *)

Коды — идентификаторы *типа смыслового блока*, *типа ошибки* или *подтипа ошибки*

Фрагмент — фрагмент исходного текста эссе, локализующий ошибку или смысловой блок

Комментарий — для ошибки непредусмотренного подтипа, без обращения к контексту сочинения

Пояснение — подробное объяснение сути ошибки для учащегося, в контексте сочинения

Исправление — предлагаемый экспертом правильный вариант фрагмента

Тег — строка из букв или цифр для связывания фрагментов, относящихся к одной и той же ошибке

РЕКОМЕНДАЦИИ ПО РАЗМЕТКЕ

1. Если в тексте присутствует смысловой блок, то его обязательно надо выделить независимо от того, насколько правильно его содержимое.
2. Если в тексте имеется ошибка, за которую эксперт собирается снизить балл по некоторому критерию, то он должен:
   - точно локализовать начало и конец ошибочного фрагмента в тексте;
   - выбрать из классификатора код подходящего типа ошибки;
   - выбрать из классификатора код подходящего подтипа ошибки; если ни один из них не подходит, то написать комментарий;
   - если суть ошибки может быть неправильно понята автором эссе, то добавить пояснение и/или исправление;
3. Если фрагмент требует связывания с другими фрагментами, то придумать уникальный тег и проставить его во всех связанных фрагментах.

ЕЩЁ НЕМНОГО О ПРИНЦИПАХ РАЗМЕТКИ

- Фрагмент, локализующий ошибку, должен быть
  - *непрерывным*, то есть разрываться на несколько фрагментов
  - *достаточным* для объяснения типа и сути ошибки
  - *безызбыточным*, то есть не захватывать лишние участки текста

- Разметка должна минимизировать вероятность апелляции
  - если есть сомнение, что это ошибка, то лучше не отмечать её
  - если суть ошибки может быть не очевидна учащемуся, то используется комментарий, пояснение и/или исправление

ПРИМЕРЫ РАЗМЕТКИ

- **Грамматическая ошибка из сочинения по русскому языку:**
  Все удивлялись его **(* Г.упр ** силой **>> силе *)**.

- **Ошибка из мини-сочинения по обществознанию: понятие определено, но не связано с основной идеей**
  **(* ПОНЯТИЕ О.теорсвязь идея ** Деятельность – это процесс целенаправленной активности людей.
  :: Для раскрытия смысла высказывания следовало бы рассмотреть особенности процесса познания. Более общее понятие деятельности уводит в сторону от основной идеи высказывания. *)

- **Из исторического сочинения: связанные смысловые блоки «причина-следствие» и ошибка в понятии**
  (* ПРИЧИНА ** Александр III считал преобразования своего отца ошибочными **#1 *), видел в них причину убийства Александра II. **(* СЛЕДСТВИЕ ** Из-за этого и из-за воспитания Победоносцева Александром III был взят **(\и.понятие** рекреативный курс. **\) #1 *)**

ПРИМЕР РАЗМЕТКИ (ПОЛНЫЙ ТЕКСТ)

Тема: (* Подлинные достижения человека откладываются не только вне его, в тех или иных порожденных им объектах, но и в нем самом. (С.Л.Рубинштейн) *)
Предмет: обществознание
К1: 1 К2: 2 К3: 1 К4: 1

(*ИдеяСмысл данного афоризма заключается в том, что результаты деятельности индивида видны не только в тех объектах, на которые эта деятельность была направлена, но и в самом человеке. Автор утверждает, что главными результатами и достижениями человека являются внутренние изменения, которые с ним происходят в процессе социализации. Рубинштейн освещает актуальную тему самовоспитания, становления личности и её социализации. Я полностью согласна с мнением автора. Настоящие достижения закладываются в процессе деятельности внутри самого человека.**)**

Обратимся к теории для подтверждения. **(*Логика** От рождения человек является **(\Понятие** индивидом – одиночным представителем человеческого вида homo sapiens **\)**. В процессе **(\Понятие** социализации - усвоения норм и ценностей, принятых в обществе **\)** – индивид становится (\Понятие\ личностью, которую можно определять как человека, как обладателя социальных черт общества, необходимых ему для жизнедеятельности в обществе **\)**. Процесс социализации длится всю жизнь. Каждое новое достижение помогает человеку укрепить свой статус личности, даёт ему толчок для дальнейших достижений и открытий. Кроме того, у человека постепенно формируется «Я-концепция», помогающая ему найти своё предназначение и занять своё место в обществе. **(\Понятие** Деятельность – это процесс целенаправленной активности людей. **\)** К видам деятельности относят труд, игру, общение, и др. Любая деятельность сопровождается **(\Понятие** мотивом – тем, почему человек занимает деятельностью **\)**, и **(\Понятие** целью – для чего он ей занимается **\)**. Достигнув цели, человек приобретает уверенность в себе и желание трудиться ещё больше для поддержания положительных результатов. Победы становятся стимулом к движению вперед.**)** Для поддержания этого мнения приведу несколько аргументов.

(\Пример.общ Как известно из многих СМИ, Стив Джобс был основоположником компании **(*о.факт незнач** по производству телефонов и планшетов «Apple» **>> «Apple» по производству персональных компьютеров *)**. Его идея телефонов нового поколения стала настоящим открытием и принесла создателю много прибыли и успеха. **(*О.подробн** Однако такой успех не расслабил Джобса, а наоборот, стал стимулом :: здесь не раскрыто, как именно достижения Джобса повлияли на него самого ***)**, воодушевившим его на создание новых электронных устройств, пользующихся невероятной популярностью и по сей день. **\)**

(*Пример.ист По сюжету серии книг о Гарри Поттере, о которой нам рассказывал учитель литературы в школе, Гермиона Грейнджер с первого курса школы чародейства и волшебства являлась лучшей ученицей среди сверстников. Её достижения в большинстве учебных предметов сделали из Гермионы уверенную в себе девушку, которая поставила себе цель быть отличницей по всем предметам в течение всего семилетнего обучения. Результаты экзаменов всегда были высоки, что помогло Грейнджер сформироваться, как личности. Результаты в учёбе влияли на поведение Гермионы, она становилась веселой и уверенной в своих силах, что видели и одобряли как учителя, так и ученики. ***)**

Таким образом, Рубинштейн в своём высказывании абсолютно прав. **(*Логика** Достижения человека в какой-либо сфере деятельности сказываются на его внутренней уверенности в себе, вере в свои силы и желании двигаться в том же направлении для достижения все больших успехов.***)**

НЕСКОЛЬКО НЕОЖИДАННЫХ ИНСАЙТОВ ;)

- Чтобы оценить сочинение, в нём надо выделять не только ошибки (минусы), но и смысловые блоки (плюсы).
- Разметка структурирует мышление у проверяющих и уменьшает число разногласий между ними благодаря необходимости выделять границы фрагментов и указывать тип ошибки.
- Даже очень простой (* язык \ разметки \ текста :: может стать >> кошмаром *) для гуманитария. Вначале разметчики испытывают трудности с закрывающими скобками, вложенными фрагментами и тегами.
- Оценка за сочинение однозначно вычисляется по разметке для всех предметов ЕГЭ… кроме английского языка.

ПРОБЛЕМА

Как учесть все атрибуты фрагментов при сравнении разметок?

МЕТРИКИ СРАВНЕНИЯ ДВУХ РАЗМЕТОК

Сравнение двух разметок одного того же эссе (алгоритмической и экспертной, либо двух экспертных)

- Метрики *парной точности* по двум разметкам X, Y одного эссе (0%--100%):
  - M_1 = точность предсказания оценки за эссе
  - M_2 = точность соответствия фрагментов
  - M_3 = точность типов ошибок и смысловых блоков
  - M_4 = точность подтипов ошибок и комментариев
  - M_5 = точность локализации фрагментов
  - M_6 = точность исправлений ошибок
  - M_7 = точность пояснений (экспертная оценка)

- Итоговая метрика *парной точности* (0%--100%) разметок X, Y — взвешенное среднее (весовые коэффициенты w_i являются параметром методики и могут быть изменены):

$$M(X,Y) = \frac{\sum_{i=1}^{7} w_i M_i(X,Y)}{\sum_{i=1}^{7} w_i}$$

МЕТРИКА M_1
точность предсказания оценки за эссе

Пусть $K(X)$ и $K(Y)$ – две оценки за эссе в баллах, вычисленные по разметкам *X* и *Y* соответственно. Тогда *точность предсказания оценки* за эссе

$$M_1(X,Y) = \left(1 - \frac{|K(X) - K(Y)|}{\max K}\right) \cdot 100\%$$

где $\max K$ – максимальная возможная оценка за эссе данного типа.

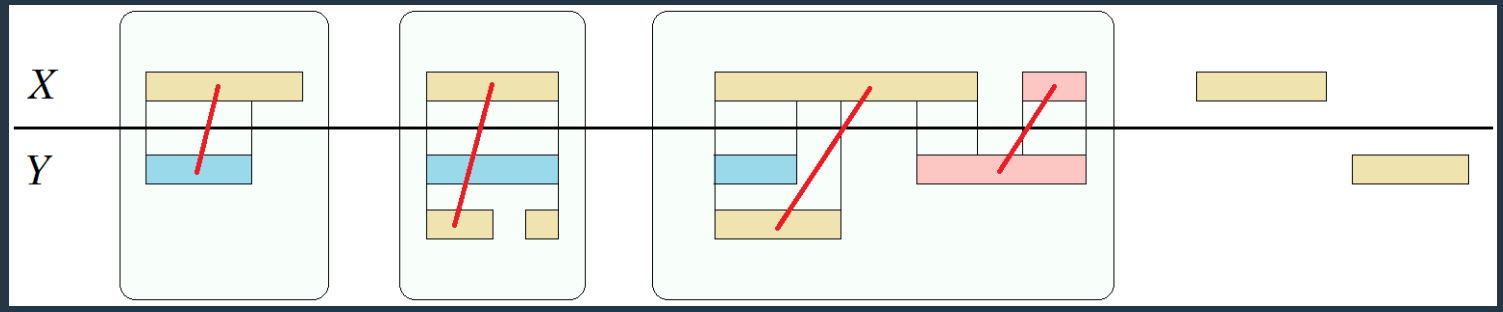
Методика вычисления оценок $K(X)$ и максимальная оценка *K* зависит от типа эссе.

В Приложении 3 к Техническому Регламенту приведены формулы для вычисления оценки в баллах за эссе по разметке, для различных типов эссе.

МЕТРИКА M_2
точность соответствия фрагментов

Разметка — это последовательность n фрагментов $X = \{x_i = (B_i, E_i, C_i)\}_{i=1}^{n}$,
где B_i – позиция начала фрагмента, E_i – позиция конца фрагмента, C_i – тип фрагмента.

Соответствие двух разметок $X = \{x_i = (B_i, E_i, C_i)\}_{i=1}^{n}$ и $Y = \{y_k = (B_k, E_k, C_k)\}_{k=1}^{m}$
это множество пар фрагментов (i, k), такое, что каждому x_i из X соответствует не более одного y_k и каждому y_k из Y соответствует не более одного x_i.
Если для фрагмента x_i не нашлось соответствия, будем записывать $[x_i \to \emptyset]$.



ПРОБЛЕМА

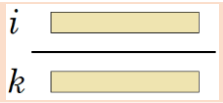
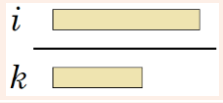
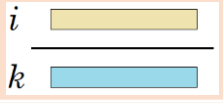
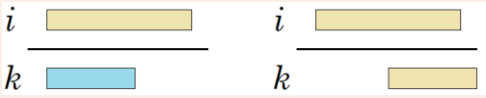
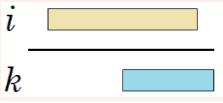
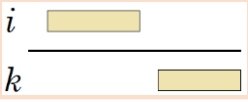
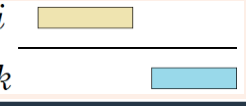
Как правильно сопоставить фрагменты в двух разметках, если они перекрываются или имеют разные типы?

МЕТРИКА M_2
точность соответствия фрагментов

Алгоритм поиска соответствия между разметками минимизирует сумму потерь Q:

$L[i,k]$ — потери от несоответствия фрагментов x_i и y_k

$$Q = \sum_{(i,k)} L[i,k] + \sum_i [x_i \to \emptyset] + \sum_k [y_k \to \emptyset] \to \min$$

$L[i,k]$	Ситуации соответствия пары фрагментов (x_i, y_k)	
0	фрагменты совпадают и имеют одинаковый тип	
0..1	фрагменты пересекаются, имеют общее начало и одинаковый тип	
1	фрагменты совпадают, но имеют разные типы	
1..2	фрагменты пересекаются и имеют либо разное начало, либо разные типы	
2	**Ситуации, когда фрагменты не образуют соответствия**	
2..3	фрагменты пересекаются, имеют разное начало и разные типы	
3	фрагменты не пересекаются и имеют одинаковый тип	
4	фрагменты не пересекаются и имеют разные типы	

МЕТРИКА M_2
точность соответствия фрагментов

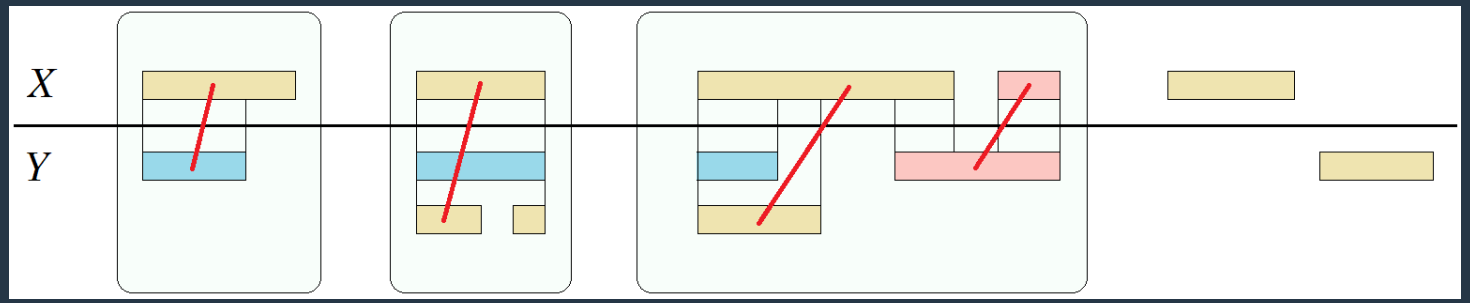
Точность — доля фрагментов разметки *X*, имеющих сопоставленный фрагмент в *Y*.
Полнота — доля фрагментов разметки *Y*, имеющих сопоставленный фрагмент в *X*.

M_2 — F_1 -мера, агрегированная метрика, определяемая как их гармоническое среднее:

$$F_1 = \frac{2 \cdot \text{точность} \cdot \text{полнота}}{\text{точность} + \text{полнота}}$$

Пример

Точность = 4/5 = 0.8
Полнота = 4/8 = 0.5
$M_2 = F_1$ = 0.615

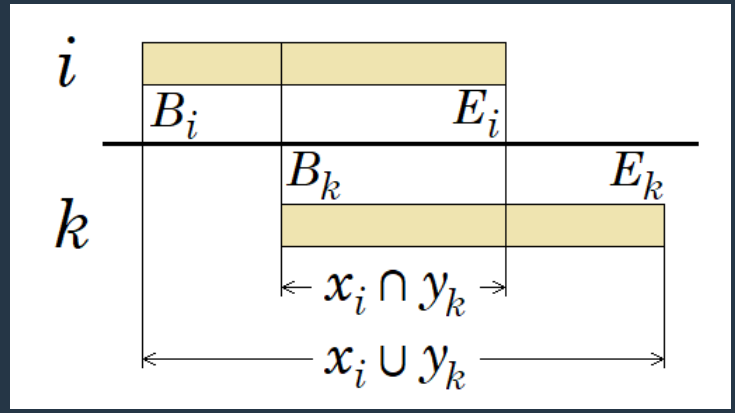


МЕТРИКА M_5
точность локализации фрагментов

Средняя точность совпадения фрагментов разметки *X* с сопоставленными им фрагментами разметки *Y*.

Точность совпадения двух фрагментов вычисляется как мера Жаккара J_{ik} — отношение числа слов в пересечении к числу слов в объединении двух фрагментов x_i и y_k.

$$J_{ik} = 1 - \frac{|x_i \cap y_k|}{|x_i \cup y_k|}$$



Если фрагмент разметки *X* не имеет сопоставленного ему фрагмента разметки *Y*, то точность их совпадения принимается равной 0%.

МЕТРИКИ M_3 M_4 M_6

M_3 — точность типов ошибок и смысловых блоков

Доля фрагментов разметки *X*, имеющих сопоставленный фрагмент в разметке *Y* с равным кодом фрагмента (типом ошибки или смыслового блока)

M_4 — точность подтипов ошибок и комментариев

Доля фрагментов разметки X, имеющих сопоставленный фрагмент в разметке Y с равным подтипом ошибки или с комментарием, который является парафразом.

M_6 — точность исправлений ошибок

Доля фрагментов разметки *X*, имеющих сопоставленный фрагмент в разметке *Y* с совпадающим исправлением.

ПРОБЛЕМА

Как оценивать пояснения?
Это ведь текст в произвольной форме, с помощью которого учитель решает свою педагогическую задачу

МЕТРИКА M_7 точность пояснений

Метрика M_7 основана не на сравнении разметок, а на оценках экспертов. Средняя экспертная оценка по всем фрагментам разметки с пояснениями.

Эксперты оценивают каждое пояснение в алгоритмической разметке от 0 до 5 баллов. Суммарный балл складывается из ответов эксперта на 5 вопросов (да=1, нет=0):

1. пояснение скорее всего будет понятно автору эссе
2. пояснение правильно объясняет суть ошибки
3. пояснение не оставляет возможностей для апелляции
4. пояснение обращается к тексту работы и конкретно к выделенному фрагменту
5. пояснение решает педагогическую задачу, помогает не делать подобных ошибок

Для получения точности в процентах число баллов умножается на 20%.

Пояснение в экспертной разметке автоматически получает 100%.

КРИТЕРИИ ТОЧНОСТИ РАЗМЕТКИ

Точность разметки оценивается по выборке размеченных эссе

- СТАР = Средняя Точность Алгоритмической Разметки
  при сравнении Алгоритмических разметок с Экспертными $M(\text{А}, \text{Э})$
- СТЭР = Средняя Точность Экспертной Разметки
  при сравнении Экспертных разметок друг с другом $M(\text{Э}, \text{Э}')$
- Относительная Точность Алгоритмической Разметки:
  ОТАР = СТАР / СТЭР * 100%
- **ОТАР ≥ 100% — условие преодоления Технологического барьера**

ПРОБЛЕМА

Не слишком ли жёстко мы задали технологический барьер?

ОТАР: СРЕДНИЙ И ОПТИМИСТИЧНЫЙ

ОТАР как *средняя относительная парная точность* алгоритмической разметки эссе, при сравнении со всем множеством {Э} экспертных разметок данного эссе:

$$M_{\text{ср}}\left(\text{А}, \{\text{Э}\}\right) = \frac{\underset{\text{Э}}{\text{avr}}\, M\left(\text{А}, \text{Э}\right)}{\underset{\text{Э}, \text{Э}'}{\text{avr}}\, M\left(\text{Э}, \text{Э}'\right)} 100\%$$

ОТАР как *оптимистичная относительная парная точность* алгоритмической разметки эссе, при сравнении со всем множеством {Э} экспертных разметок данного эссе:

$$M_{\text{опт}}\left(\text{А}, \{\text{Э}\}\right) = \frac{\max\limits_{\text{Э}} M\left(\text{А}, \text{Э}\right)}{\min\limits_{\text{Э}, \text{Э}'} M\left(\text{Э}, \text{Э}'\right)} 100\%$$

Идея: усредним эти две оценки с весовым коэффициентом — *параметром жёсткости* H

КРИТЕРИЙ ТОЧНОСТИ РАЗМЕТКИ с параметром жёсткости критерия

СТАР — Средняя Точность Алгоритмической Разметки, с учётом параметра жёсткости H:

$$\text{СТАР} = \underset{\text{эссе}}{\text{avr}} \left(H \underset{\text{Э}}{\text{avr}}\, M\left(\text{А}, \text{Э}\right) + \left(1 - H\right) \underset{\text{Э}}{\max}\, M\left(\text{А}, \text{Э}\right) \right)$$

СТЭР — Средняя Точность Экспертной Разметки, с учётом параметра жёсткости H:

$$\text{СТЭР} = \underset{\text{эссе}}{\text{avr}} \left(H \underset{\text{Э}, \text{Э}'}{\text{avr}}\, M\left(\text{Э}, \text{Э}'\right) + \left(1 - H\right) \underset{\text{Э}, \text{Э}'}{\min}\, M\left(\text{Э}, \text{Э}'\right) \right)$$

ОТАР — Относительная Точность Алгоритмической Разметки:

ОТАР = СТАР / СТЭР * 100%

Параметр жёсткости критерия H принимает значения от 0 до 1.
Чем меньше H, тем выше и оптимистичнее оценка ОТАР.

Контент-анализ: обобщение и автоматизация

Обобщённый контент-анализ — четыре базовые операции с текстом:

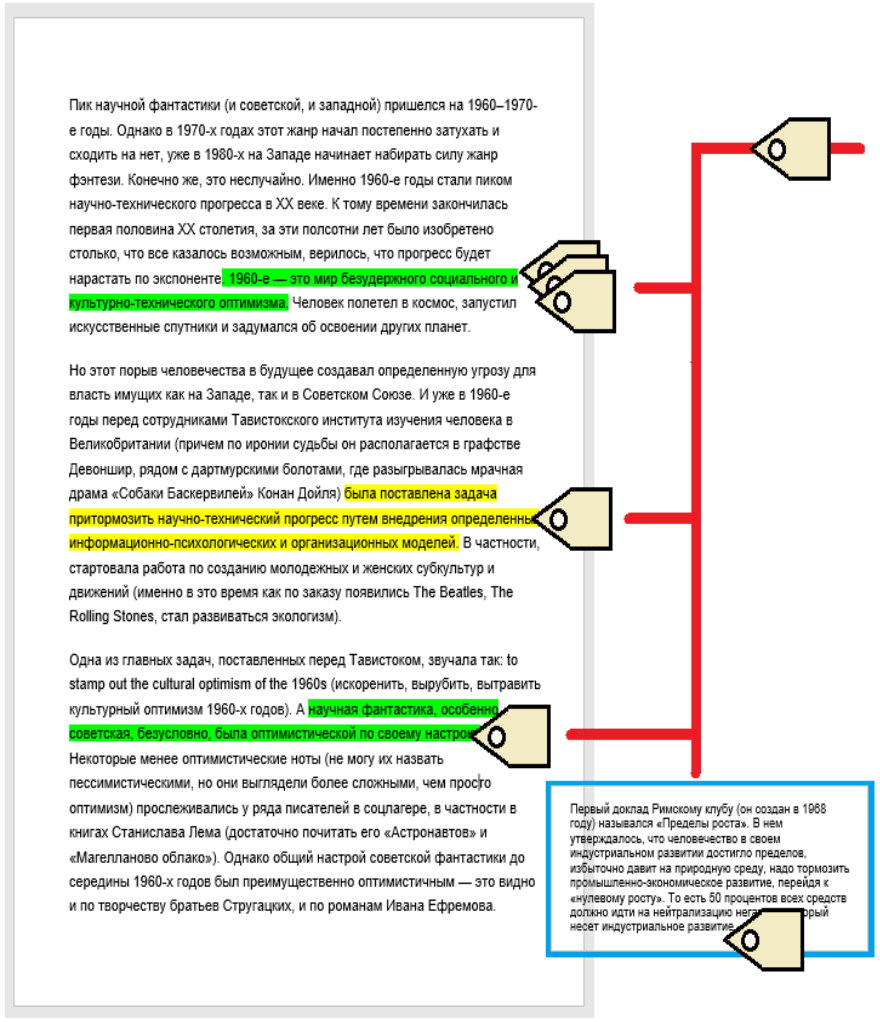
1) выделить фрагмент
2) классифицировать (тегировать) фрагмент по рубрикатору
3) связать несколько фрагментов
4) дать комментарий (затекст) к фрагменту или связи

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

Три задачи построения обучаемой модели разметки:

1) разработка рубрикатора, инструкций разметчика, организация разметки
2) выбор большой языковой модели и её (до)обучение по разметке
3) оценивание качества разметки, сравнение и выбор моделей

Разметка текста: обобщённый контент-анализ

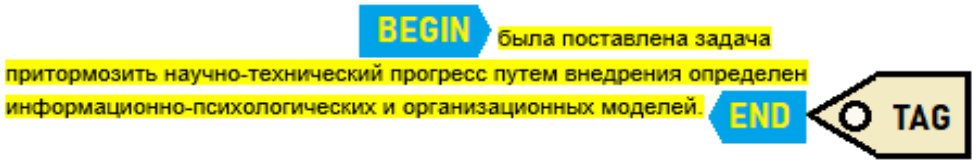


Разметка состоит из элементов

Элемент разметки — несколько взаимосвязанных фрагментов, затекстов и тегов

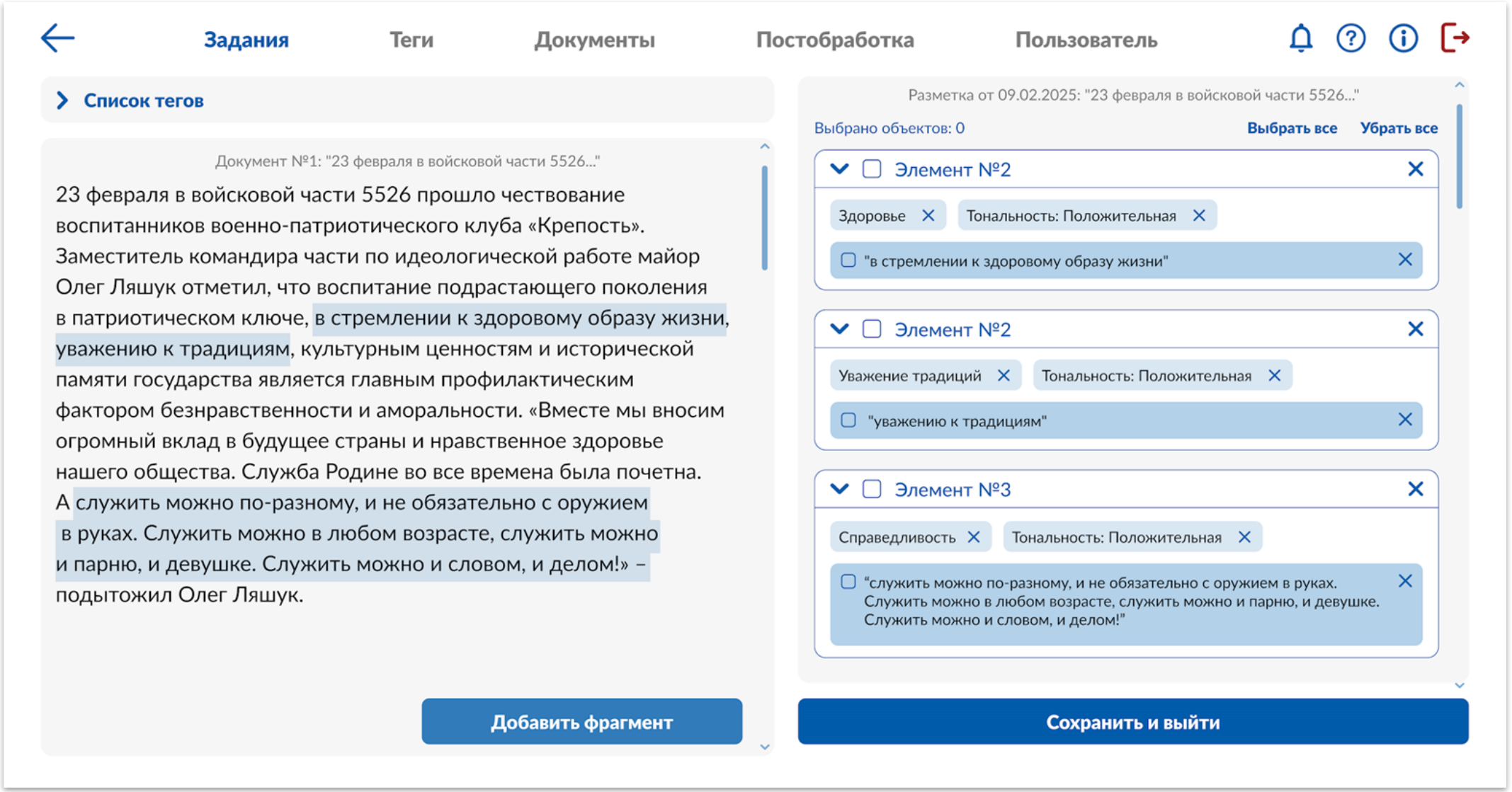
Теги (классы) выбираются из рубрикатора

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

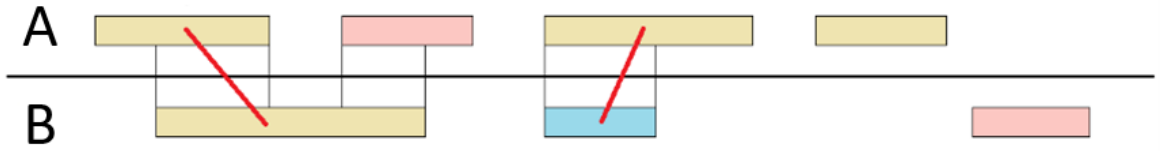


Методика оценивания алгоритмической разметки

- В основе методики — парное сравнение разметок текста:
  «алгоритм ↔ эксперт», «эксперт-1 ↔ эксперт-2»
  на основе оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок Con1,…,5(A,B)
- 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(E1,E2) разметок двух экспертов, E1 и E2
- 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):

Con1 = доля фрагментов, для которых найдено сопоставление

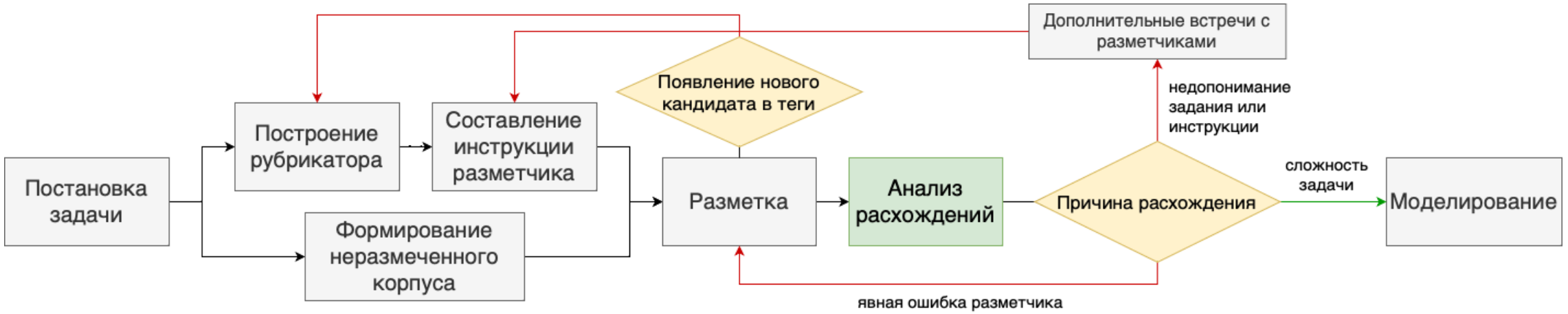
Con2 = точность наложения сопоставленных фрагментов

Con3 = точность совпадения тегов сопоставленных фрагментов

Con4 = точность совпадения связей сопоставленных фрагментов

Con5 = точность совпадения затекстов сопоставленных фрагментов

Организация процесса разметки



- каждый документ размечается несколькими экспертами (2 или 3)
- документы ранжируются по согласованности экспертов Con(E,E’)
- наибольшие расхождения обсуждаются, вырабатывается консенсус
- происходит доработка инструкции и/или переразметка документов

Детекция ценностей социокультурного кода

Создание рубрикатора на основе кросс-дисциплинарного подхода

Исследователи	Предмет исследований
Милтон Рокич	Ценностные ориентации людей - психология, социология
Герт Хоффстеде	Культурные характеристики народов - социология
Шолом Шварц	Теория базовых человеческих ценностей - социальная психология
Рональд Инглхарт	Исследование мировых ценностей - политология, социология
Сэмюэл Хантингтон	Этнокультурное описание цивилизаций - политология, социология
Юрий Сергеевич Степанов	Концепты русской культуры - лингвистика
Александр Александрович Аузан	Культурные коды экономики - экономика

Указ президента Российской Федерации № 400 от 2-07-2021 «О стратегии нацбезопасности»
Указ президента Российской Федерации № 809 от 9-11-2022 «Об утверждении основ госполитики по сохранению и укреплению традиционных российских духовно-нравственных ценностей»

Какие ценности брать для рубрикатора: «аксиоматический» подход

- **Общественная значимость**
  ценность — это то, что оказывает влияние на социальную жизнь
- **Индивидуальная значимость**
  то, что влияет на принятие решений отдельными людьми
- **Субъективная измеримость**
  то, что человек может принимать, отвергать или быть безразличным
- **Коммуникативность**
  то, на отношение к чему можно повлиять в процессе коммуникации
- **Текстуальность**
  то, что возможно описать, выразить текстом, фразой, историей
- **Атомарность**
  то, что не сводится к набору других ценностей

Рубрикатор ценностей (1 из 2)

М. Рокич	Ш. Шварц	Г. Хофстеде
Ю.С. Степанов	Р. Инглхарт	Б.С. Ерасов

Группа социальных ценностей

- Социальные ценности
- Авторитет
- Альтруизм
- Благородство происхождения / аристократизм
- Важность общественного мнения
- Воспитание
- Гендерное разнообразие
- Дети
- Долгосрочная ориентация
- Дружба
- Избегание неопределённости
- Индивидуализм
- Интеллигентность
- Коллективизм
- Культура (нормы) поведения
- Личные границы
- Материальные ценности
- Патриархальность
- Патриотизм
- Пацифизм / мир во всём мире
- Полезность (созидательный труд)
- Профессиональный успех
- Репутация
- Семья
- Социальное признание
- Суеверия
- Трудолюбие / продуктивность
- Чувство принадлежности / единство народов
- Этничность
- Язык

Группа витальных ценностей

- Витальные (необходимые) ценности
- Безопасность (личная)
- Время
- Еда
- Жизнь
- Жилище
- Здоровье
- Природа

Группа политических ценностей

- Политические ценности
- Власть
- Выборность власти (демократия)
- Институциональное доверие
- Историческая память и преемственность поколений
- Либерализм
- Национальная безопасность
- Права и свободы
- Правосознание (гражданская активность, гражданственность)
- Справедливость

Рубрикатор ценностей (2 из 2)

М. Рокич	Ш. Шварц	Г. Хофстеде
Ю.С. Степанов	Р. Инглхарт	Б.С. Ерасов

Группа религиозных ценностей

- Религиозные ценности
- Благочестивость
- Бог
- Религиозность
- Эзотерика

Группа эстетических и гедонистических ценностей

- Эстетические и гедонистические ценности
- Жизнь, полная впечатлений
- Красота
- Культура и искусство
- Наслаждение жизнью
- Потворство желаниям
- Творчество
- Эстетика

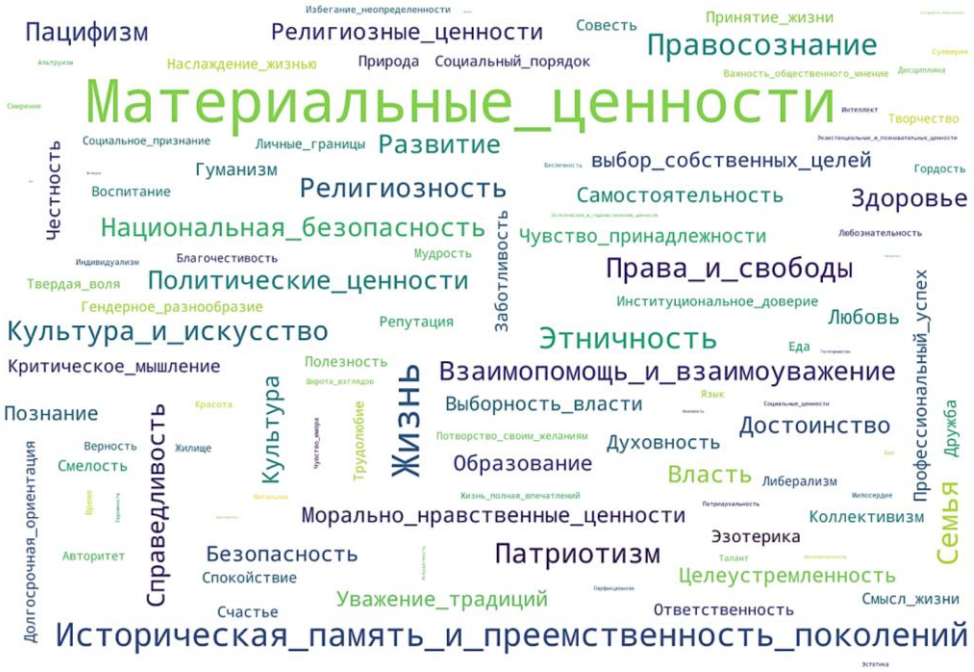
Группа экзистенциальных и познавательных ценностей

- Экзистенциальные и познавательные ценности
- Интеллект
- Критическое мышление
- Любовь
- Любознательность
- Мудрость
- Образование
- Перфекционизм
- Познание
- Принятие жизни
- Развитие
- Самостоятельность (выбор собственных целей)
- Смелость
- Смысл жизни
- Спокойствие (внутренняя гармония)
- Счастье
- Талант
- Твёрдая воля
- Целеустремлённость
- Широта взглядов

Группа морально-нравственных ценностей

- Морально-нравственные ценности
- Аккуратность
- Беспечность
- Вежливость
- Верность
- Взаимопомощь и взаимоуважение
- Гордость
- Гостеприимство
- Гуманизм
- Дисциплина
- Достоинство (самоуважение, самоценность)
- Духовность (приоритет духовного над материальным)
- Заботливость
- Искренность
- Милосердие
- Ответственность
- Скромность
- Смирение (послушание, кротость)
- Совесть (нравственный закон, мораль)
- Терпение
- Уважение традиций
- Целомудрие
- Честность
- Чувство юмора

Ценностный ландшафт



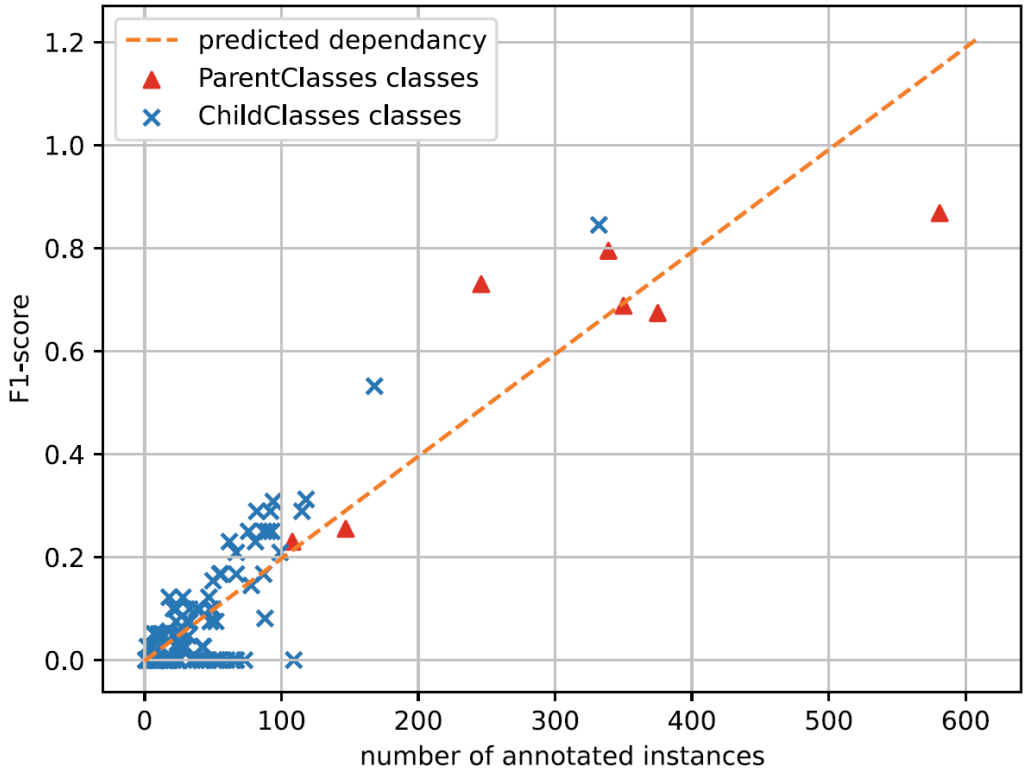
Самые частотные теги (по количеству элементов)

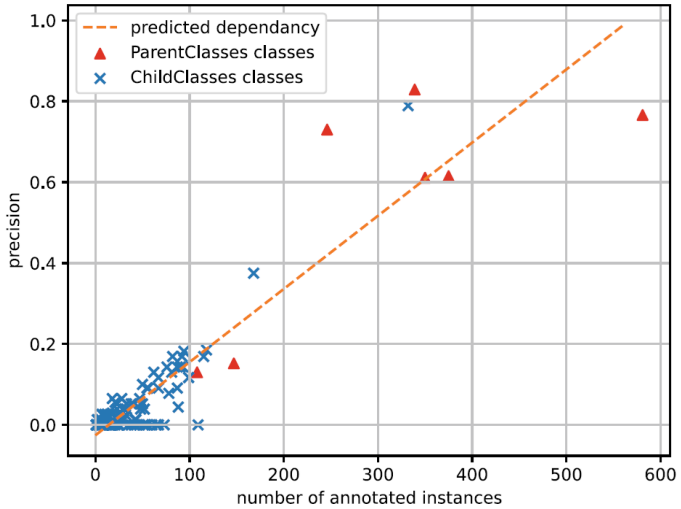
40%

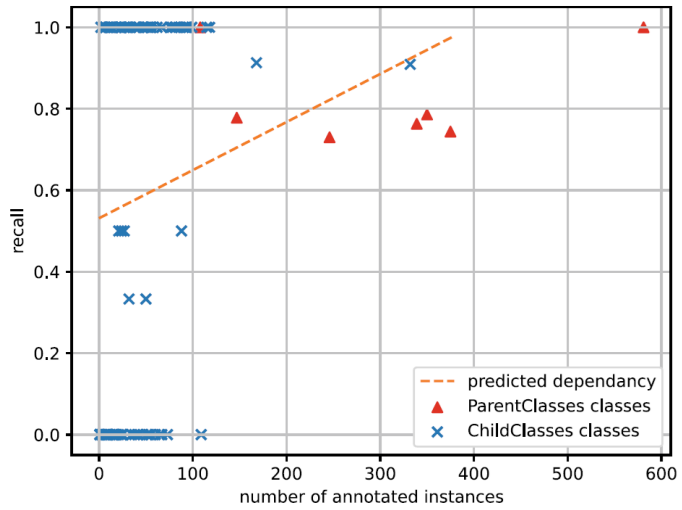
1	Материальные ценности	1569
2	Жизнь	478
3	Историческая память и преемственность поколений	329
4	Правосознание (гражданская активность, гражданственность)	267
5	Политические ценности	246
6	Семья	243
7	Этничность	237
8	Культура и искусство	236
9	Права и свободы	235
10	Патриотизм	233

1. *Rink Olga, Lobachev Viktor, Vorontsov Konstantin.* Detecting human values and sentiments in large text collections with a context-dependent information markup: a methodology and math. HCII 2024. Lecture Notes in Computer Science series (in print). Cham: Springer.
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

Зависимость точности, полноты и F1 от объёма выборки, по классам







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.

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.

Бенчмарк для универсальных моделей разметки

Состоит из датасетов с задачами разметки текстов с максимально широким покрытием различных доменов.

Все датасеты приведены к универсальному формату данных разметки для демонстрации универсальности подхода.

Бенчмарк содержит:

- 21 датасет
- 17 типов задач

Мультиасессорная разметка встречается в некоторых датасетах, однако **авторы пренебрегают решением проблемы противоречивости разметок** и используют стандартные для области критерии оценивания качества.

Датасеты для бенчмарка универсальных моделей разметки

Dataset	Task Type	Level	Paper Count
Kaggle NER Corpus (Bos et al., 2017)	NER	Span Level-1	25
	POS Tagging	Span Level-1	5
MultiCoNER (Malmasi et al., 2022)	NER	Span Level-1	85
RuTermEval Dialogue (Mamontova et al., 2025)	NER	Span Level-1	8
SWDA (Stolcke et al., 2000)	Dialogue Act	Span Level-2	228
RuSentNE (Golubev et al., 2023)	NER	Span Level-1	5
	Semantic analysis	Span Level-2	5
	Opinion tuple extraction	Element Level	2
ADE (Gurulingappa et al., 2012)	Text classification	Span Level-1	10
	NER	Span Level-1	150
	RE	Element Level	43
	Coreference resolution	Element Level	1
DDI corpus (Herrero-Zazo et al., 2013)	NER	Span Level-1	102
	RE	Element Level	131
PcMSP (Yang et al., 2022)	Text classification	Span Level-1	1
	NER	Span Level-1	5
	RE	Element Level	5
ChemProt (Krallinger M., 2017)	Text classification	Span Level-1	1
	NER	Span Level-1	3
	RE	Element Level	70
NERRE (Dagdelen et al., 2024)	NER	Span Level-1	10
	RE	Element Level	11
NEREL (Loukachevitch et al., 2023)	NER	Span Level-1	9
	RE	Element Level	2
	Extraction of multi-spans with named entities	Element Level	1
RURED (Gordeev et al., 2020)	NER	Span Level-1	3
	RE	Element Level	6
	Entity linking	Element Level	1
Scierc (Luan et al., 2018)	NER	Span Level-1	137
	RE	Element Level	134
	Coreference resolution	Element Level	113
CONLL 2012 Ontonotes (Pradhan et al., 2013)	NER	Span Level-1	87
	Semantic Role labeling	Element Level	25
	Coreference resolution	Element Level	15
RuSuperGLUE's RWSD task (Shavrina et al., 2020)	Relation classification	Element Level	7
MERA's RWSD task (Fenogenova et al., 2024)	Relation classification	Element Level	5
MERA's Ruethics task (Pradhan et al., 2013)	Relation classification	Element Level	1
SemEval 2010 Task 8 (Hendrickx et al., 2019)	Relation classification	Element Level	8
SemEval-2018 Task 7 (Buscaldi et al., 2017)	NER	Span Level-1	14
	RE	Element Level	57
	Relation classification	Element Level	57
	Knowledge Graph Construction	Element Level	3
UpGreat READ//ABLE (READ/ABLE, 2017-2025)	Extraction and classification of spans with errors	Span Level-2	2
	Extraction of multi-spans with errors	Element Level	1
	Annotation of text spans with comments	Extended Level	1
Human Values dataset (Rink et al., 2024)	Text classification	Span Level-2	1
	Extraction of spans with human values	Span Level-2	2
	Extraction of elements with human values	Element Level	2
	Semantic analysis of elements with human values	Element Level	2
	Annotation of text markups with comments	Extended Level	2

Kaggle NER Corpus: данные и критерии качества

Датасет содержит новостные, исторические, политические тексты на английском языке. Датасет фактически содержит 2 задачи:

- NER (Named Entity Recognition — выделение именованных сущностей)
- POS Tagging (Part-Of-Speech Tagging — маркировка частей речи)

Для обеих задач используются классические критерии качества:

- Accuracy, Precision, Recall — для задачи NER
- F1 — для обеих задач

Датасет содержит несогласованную мультиасессорную разметку, однако авторы не заявляют об этом явно.

Conll 2012 Ontonotes: данные и критерии качества

Тексты новостей, телепередач, журналов, религиозных текстов, форумов и блогов на 3 языках: арабском, китайском и английском. Мы выделили 7 задач 2 типов:

- NER (распознавание именованных сущностей)
- RE (Relation Extraction — выделение связей)

Для NER используются классические критерии: Accuracy, Precision, Recall, F1, ECE.

Для RE предлагаются продвинутые критерии, учитывающие согласованность между разметками: MUC, B-CUBED, CEAF4, Alignment Rate.

Датасет содержит несогласованную мультиасессорную разметку, однако авторы не заявляют об этом явно.

Предлагаемые критерии качества

Пример ConlI 2012 Ontonotes показывает, что для отдельных типов задач были попытки решить проблему несогласованности, однако еще никто не пробовал

- предложить многокритериальное оценивание, разделяя аспекты согласованности
- обобщить критерии для разных задач;

Большинство авторов либо вовсе игнорируют эту проблему, либо “решают” ее использованием одной из разметок для каждого текста.

Для решения проблемы несогласованности асессорских разметок мы предлагаем систему критериев качества, предполагающую, что все асессоры правы.

Обобщенный критерий (согласованность)

$$C(M_1, M_2) = \sum_{i=1}^{5} C_i(M_1, M_2),$$

где M_1, M_2 - разметки.
Поскольку фрагменты в первой и второй разметке не полностью совпадают, требуется найти оптимальное соответствие фрагментов.
Согласованность зависит ещё от соответствия фрагментов A:
$C(M_1, M_2) \equiv C(M_1, M_2; A)$.
Оптимальным называется соответствие, при котором достигается максимум согласованности:

$$A^* = \{(i,k) : F_i^1 \longleftrightarrow F_k^2\} : C(M_1, M_2; A^*) = \max_A C(M_1, M_2; A)$$

Критерий 1
(доля верно сопоставленных фрагментов)

$$C_1(M_1, M_2; A) = \frac{2|A|}{n_1 + n_2},$$

где n_1, n_2 - количества фрагментов в разметках M_1, M_2 соответственно.

Критерий 2
(пересечение текстов в сопоставленных фрагментах)

$$C_2(M_1, M_2; A) = \frac{1}{|A|} \sum_{(i,k)\in A} \frac{2|[b_i, e_i] \cap [b_k, e_k]|}{|[b_i, e_i]| + |[b_k, e_k]|},$$

где $[b_i, e_i], [b_k, e_k]$ - отрезки индексов выделенного текста фрагментов F_i^1 и F_k^2 соответственно.

Критерий 3 (пересечение множеств тегов сопоставленных фрагментов)

$$C_3(M_1, M_2; A) = \frac{1}{|A|} \sum_{(i,k)\in A} \frac{|T_i \cap T_k|}{|T_i| + |T_k|},$$

где T_i, T_k - множества тегов фрагментов F_i^1, F_k^2 соответственно.

Критерий 4
(доля совпавших отношений)

Критерий 4 представляет собой дробь: в числителе - удвоенное количество пар отношений (relations), где фрагменты полностью соотносятся по соответствию A, в знаменателе - сумма количеств отношений в 1-й и 2-й разметке:

$$C_4(M_1, M_2; A) = \frac{2|G|}{m_1 + m_2},$$

где $G = \{(i,k) : R_i^1 \sim R_k^2\}$, считаем, что $R_i^1 \sim R_k^2$, если $\forall F_j^1 \in R_i^1 \quad \exists F_l^2 \in R_k^2 : (j,l) \in A$. Отношения, для которых $R_i^1 \sim R_k^2$, будем называть совпавшими отношениями.

Критерий 5
(пересечение тегов совпавших отношений)

$$C_5(M_1, M_2; A) = \frac{1}{|G|} \sum_{(i,k)\in G} \frac{|T_i \cap T_k|}{|T_i| + |T_k|},$$

где T_i, T_k - множества тегов отношений R_i^1, R_k^2 соответственно.

Выводы по методике оценивания

Предложенные критерии учитывают несогласованность мультиасессорной разметки текстов при сравнении разметок асессоров и модели машинного обучения.

Это позволяет использовать датасеты с несогласованными мультиасессорными разметками для обучения и валидации моделей.

Для датасетов с одной асессорской разметкой предложенные нами критерии ведут себя как классические метрики, оценивая близость разметки к Ground Truth.

Выводы

- Большие языковые модели позволяют сегодня решать те задачи, которые ещё 5 лет назад считались непреодолимо трудными
- В том числе автоматизировать контент-анализ больших текстовых корпусов для масштабирования социогуманитарных исследований
- **Ближайшая задача:** составление мультизадачного бенчмарка для тестирования универсальности обучаемых моделей разметки

Воронцов Константин Вячеславович
д.ф.-м.н., профессор РАН,
зав. лабораторией МОСА Института ИИ МГУ,
зав. кафедрой ММП ВМК МГУ,
voron@mlsa-iai.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov