Аннотации, вводные и заключительные разделы научных статей и их ранжирование по близости смысловому эталону

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

15-я Международная конференция
«Интеллектуализация обработки информации» (ИОИ-2024),

23–27 сентября 2024 г.

Республика Беларусь, г. Гродно

Формирование индивидуальной образовательной траектории студента

Составление подборки публикаций по заданной теме:

- анализ релевантности словаря каждой публикации интересующей пользователя теме;
- учёт конечной цели пользователя — для решения каких именно задач делается подборка.

Подготовка электронного учебного материала:

- поиск оптимального порядка работы с первоисточниками от более общего к более специфическому;
- идеальный случай — оценка взаимной смысловой зависимости текстов относительно наиболее рациональных (эталонных) вариантов описания представляемых ими фрагментов знаний.

«Эталонному» варианту здесь отвечают публикации,для которых

при *максимально полном* раскрытии интересующей пользователя темы характерен максимум среднего числа *наиболее значимых терминов* в расчёте на одно простое распространённое предложение (фразу) при минимуме его длины (в словах).

Языковые модели семейства BERT

- основаны на архитектуре Transformer;
- предварительно обучаются на больших текстовых коллекциях;
- с помощью указанных моделей предложения отображаются в многомерные векторы («эмбеддинги»);
- из известных моделей BERT наибольший интерес здесь представляют модели типа SciBERT, обучаемые на корпусах научных текстов.

Эмбеддинги (англ. *embeddings*)

- каждый такой вектор показывает встречаемость заданного предложения в определённом контексте;
- возможно их построение для произвольного законченного текстового фрагмента (слова, параграфа и т. п.);
- для анализируемых текстовых фрагментов оценка их смысловой близости (т. е. «силы» смысловой связи) может быть формально определена через меру близости соответствующих им векторов.

[1] от англ. Bidirectional Encoder Representations from Transformers

По каждому предложению Ts_j аннотации $\mathbb{T}\mathrm{s}_i$ для отвечающего ему эмбеддинга вычисляется массив значений $\mathbb{C}\mathrm{s}_j$ косинусной близости аналогичным векторам других предложений аннотации и выбирается предложение $Ts_{\max}$ с максимальным суммарным значением близости до остальных предложений.

Назовём далее $Ts_{\max}$ центром масс $\mathbb{T}\mathrm{s}_i$ относительно смысловой связности.

Основные идеи решения (текущее состояние вопроса)

- «точкой входа» в формируемой траектории работы пользователя с первоисточниками будет та публикация, которая максимально связана по смыслу с остальными работами ранжируемой коллекции;
- среднеквадратическое отклонение оценки «силы» смысловой связи должно быть минимальным;
- анализируемыми фрагментами публикаций являются их аннотации вместе с заголовками как отражающие основное содержание каждой из работ и наиболее значимые результаты без излишних деталей;
- для «силы» смысловой связи публикации с другими работами коллекции вводятся две независимые оценки: для полных текстов аннотаций публикаций и для центров масс аннотаций.

Результирующий рейтинг публикации и близость её аннотации эталону

Смысловую связность аннотации $\mathbb{T}\mathrm{s}_i$ можно формально определить как

$$\mathrm{cn}\left(\mathbb{T}\mathrm{s}_i\right) = \frac{\max\left(\mathbb{C}\mathrm{s}_{\max}\right)}{\left(1.0 + \mathrm{std}\left(\mathbb{C}\mathrm{s}_{\max}\right)\right)}, \tag{1}$$

где $\mathrm{std}\left(\mathbb{C}\mathrm{s}_{\max}\right)$ — СКО значения косинусной близости предложения $Ts_{\max}$ остальным предложениям аннотации,

$\max\left(\mathbb{C}\mathrm{s}_{\max}\right)$ — максимальное из значений в массиве $\mathbb{C}\mathrm{s}_{\max}$.

Замечания

- в случае оценки «силы» смысловой связи относительно центров масс аннотаций в роли массива $\mathbb{C}\mathrm{s}_{\max}$ будет массив значений косинусной близости вектора центра масс анализируемой аннотации аналогичным векторам центров масс аннотаций остальных публикаций коллекции;
- при оценке «силы» смысловой связи относительно полных текстов аннотаций указанный массив будет состоять из значений косинусной близости эмбеддинга для текста анализируемой аннотации и соответствующих эмбеддингов аннотаций остальных публикаций.

Утверждение 1

Результирующий рейтинг публикации, ассоциируемый с близостью её аннотации эталону, определяется произведением оценки «силы» смысловой связи публикации с остальной коллекцией и оценки смысловой связности аннотации анализируемой публикации.

Полнота изложения основного содержания работы в её аннотации

Суть предлагаемого решения проблемы полноты изложения содержания

Текст аннотации расширяется предложениями из вводного (*introduction*) и заключительного (*conclusions*) разделов анализируемой работы, при этом контролируется изменение оценки (1) для расширенной аннотации.

1. Вычисляется значение оценки (1) для исходного (нерасширенного) варианта аннотации, это значение принимается за текущее.
2. Далее в аннотацию добавляется то предложение из объединённого множества предложений *introduction* и *conclusions*, для которого величина оценки (1) по расширенной аннотации будет максимальной.
3. Если новое значение оценки (1) больше текущего, то на следующей итерации оно становится текущим, а процесс повторяется для объединённого *introduction* и *conclusions*, из которого удаляется только что добавленное в аннотацию предложение.
4. Процесс завершается, когда на очередной итерации новое значение оценки (1) оказывается меньше текущего, а в качестве результата возвращается аннотация из предвдущей итерации.

Основная гипотеза

Степень полноты изложения основного содержания работы повышается за счёт роста смысловой связности аннотации.

Полнота изложения основного содержания работы в её аннотации

Для ранжирования относительно заданной языковой модели

- отбираются те из расширенных аннотаций, рейтинг по значению близости эталону у которых оказался не ниже, чем для исходных вариантов тех же аннотаций, причём как по центрам масс, так и по полным текстам аннотаций;
- аннотации остальных работ коллекции при сравнении здесь берутся в исходном (нерасширенном) варианте.

Замечание

Для формирования оптимального порядка работы пользователя с публикациями уже в ранжированной коллекции для каждой работы находится наиболее близкая ей по смыслу на основе косинусной близости соответствующих эмбеддингов. При этом траектория навигации пользователя по коллекции строится «сверху вниз» от публикации с большим рейтингом к наиболее близкой ей публикации с меньшим рейтингом.

Экспериментальные исследования

Задействованные модели[2] трансформеров предложений, работающие с русским языком:

- *mlsa-iai-msu-lab/sci-rus-tiny*;
- *bert-base-nli-mean-tokens*;
- *ai-forever/ruscibert*;
- *sentence-transformers/distiluse-base-multilingual-cased-v1*;
- *sentence-transformers/all-MiniLM-L6-v2*;
- *cointegrated/rut5-base-multitask*;
- *cointegrated/rut5-base-paraphraser*.

Помимо исходных вариантов *ruscibert* и *sci-rus-tiny*, в экспериментах участвовали варианты этих моделей, дообученные на датасетах перифраз *merionum/ru_paraphraser* (обе модели) и *cointegrated/ru-paraphrase-NMT-Leipzig* (только *ruscibert*).

Вычисление косинусной близости эмбеддингов:

- центров масс — функция *cosine_similarity* библиотеки *sklearn.metrics.pairwise*;
- для полных текстов аннотаций — аналогичная функция *pytorch_cos_sim* из библиотеки *sentence_transformers.util*.

Реализация на Python 3.10 (Jupyter Notebook, исходные данные и результаты)

[2] более подробное их описание представлено на портале *huggingface.co*

Пример: ММРО-15, Статистическая теория обучения

Таблица 1. Ранжирование исходных аннотаций, модель *ai-forever/ruscibert* (без дообучения).

N_1	Автор (ы) и заголовок статьи	N_2
1	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	4
2	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	1
3	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	2
4	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	7
5	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	3
6	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	8
7	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	6
8	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	5
9	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	9
10	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	10

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример расширения аннотации с номером $N_1 = 6$ по *Таблице 1*

Исходный вариант:

Принцип максимизации зазора для монотонного классификатора ближайшего соседа. Получены точные оценки полного скользящего контроля для монотонных классификаторов, основанных на принципе ближайшего соседа. Показано, что наилучшей обобщающей способностью обладает монотонный классификатор, в котором разделяющая поверхность проходит посередине зазора между классами. Показана связь данной задачи с задачей доопределения частично заданной монотонной булевой функции.

Вариант расширения 1 *(относительно моделей bert-base-nli-mean-tokens, all-MiniLM-L6-v2, ruscibert, sci-rus-tiny)*:

В данной работе рассматривается семейство монотонных классификаторов ближайшего соседа, обобщающее конструкцию монотонной корректирующей операции из [2].

Вариант расширения 2 *(относительно моделей rut5-base-multitask, distiluse-base-multilingual-cased-v1, rut5-base-paraphraser)*:

Для него получена точная оценка, в случае, когда исходная выборка монотонна.

Таблица 2. Ранжирование аннотаций после расширения аннотации с номером $N_1 = 6$ по *Таблице 1*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	1
2	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	2
3	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	6
4	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	4
5	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	5
6	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	3
7	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	8
8	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	7
9	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	10
10	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	9

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Таблица 3. Для сравнения: ранжирование расширенных аннотаций, модель *ai-forever/ruscibert* (без дообучения).

N_1	Автор (ы) и заголовок статьи	N_2
1	Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии	1
2	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	5
3	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	4
4	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	2
5	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	3
6	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	6
7	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	7
8	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	8
9	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	10
10	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	9

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример: ММРО-15, Статистическая теория обучения

Таблица 4. Ранжирование исходных аннотаций, модель *ai-forever/ruscibert*, последовательно дообученая на датасетах *merionum/ru_paraphraser* и *cointegrated/ru-paraphrase-NMT-Leipzig*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	4
2	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	5
3	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	1
4	Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии	2
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	8
6	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	3
7	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	6
8	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	7
9	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	9
10	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	10

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Таблица 5. Ранжирование аннотаций после расширения аннотации с номером $N_1 = 6$ по *Таблице 1*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	1
2	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	4
3	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	5
4	Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии	2
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	8
6	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	3
7	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	6
8	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	7
9	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	9
10	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	10

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример: ММРО-15, Статистическая теория обучения

Таблица 6. Ранжирование расширенных аннотаций, модель *ai-forever/ruscibert*, последовательно дообученая на датасетах *merionum/ru_paraphraser* и *cointegrated/ru-paraphrase-NMT-Leipzig*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	1
2	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	3
3	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	7
4	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	8
5	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	2
6	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	4
7	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	9
8	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	6
9	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	5
10	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	10

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Таблица 7. Ранжирование исходных аннотаций, модель *rut5-base-paraphraser*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	3
2	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	2
3	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	5
4	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	4
5	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	8
6	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	6
7	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	7
8	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	9
9	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	10
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	1

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример: ММРО-15, Статистическая теория обучения

Таблица 8. Ранжирование аннотаций после расширения аннотации с номером $N_1 = 6$ по *Таблице 1*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	1
2	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	3
3	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	4
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	2
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	5
6	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	6
7	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	7
8	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	9
9	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	8
10	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	10

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Таблица 9. Ранжирование расширенных аннотаций, модель *rut5-base-paraphraser*.

N_1	Автор (ы) и заголовок статьи	N_2
1	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	6
2	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	3
3	Каневский Д. Ю. Переобучение и комбинаторная радемахеровская сложность в задачах восстановления регрессии	9
4	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	7
5	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	5
6	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	2
7	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	10
8	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	4
9	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	8
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	1

Здесь N_1 и N_2 — порядковые номера статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Краткие итоги и открытые проблемы

Следует отметить, что

из 100 экспериментов (на десяти статьях по десяти вариантам моделей) только в пяти испытаниях рейтинг по значению близости эталону аннотации снизился после её расширения предложениями введения/заключения.

При этом актуальны и требуют отдельного исследования две проблемы:

1. Возможность удаления предложений из аннотации с заменой новыми из объединённого *introduction* и *conclusions* либо без таковой.
2. В расширенной аннотации даже при повышении рейтинга по близости эталону примерно в половине случаев новые предложения нельзя назвать логическим продолжением существующего текста, например, при наличии в тексте местоимений или ссылок на первоисточники.

Возможный вариант решения

задействовать абстрактивную суммаризацию (в нашем случае здесь имеем задачу типа *one-document*) со сравнением автоматически сгенерированной и расширенной предложенным в настоящей работе методом аннотации.