

Нейросетевые модели языка и отбор значимых фрагментов научных статей для безызбыточной передачи смысла

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

22-я Всероссийская конференция с международным участием
«Математические методы распознавания образов» (ММРО-2025),

22–26 сентября 2025 г.

Россия, г. Муром

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 отвечают публикации, для которых характерен:

- максимум среднего числа *наиболее значимых терминов* в расчёте на простое распространённое предложение при минимуме его длины;
- *максимально полное* раскрытие интересующей пользователя темы.

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.

¹от англ. Bidirectional Encoder Representations from Transformers

По каждому предложению Ts_j аннотации Ts_i для отвечающего ему эмбединга вычисляется массив значений Cs_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 Ts_i относительно смысловой связности.

«Сила» смысловой связи публикаций и траектория навигации по подборке

- «точкой входа» в формируемой траектории работы пользователя с первоисточниками будет та публикация в составе ранжируемой коллекции, которая максимально связана по смыслу с остальными работами 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 методологических деталей;
- для «силы» смысловой связи публикации с другими работами коллекции вводятся две независимые оценки: для полных текстов аннотаций публикаций и для центров масс аннотаций;
- степень полноты изложения основного содержания работы в её аннотации может быть увеличена путём повышения смысловой связности последней.

Смысловую связность аннотации Ts_i можно формально определить как

$$cn(Ts_i) = \frac{\max(Cs_{\max})}{(1.0 + \text{std}(Cs_{\max}))}, \quad (1)$$

где $\text{std}(Cs_{\max})$ — СКО значения косинусной близости предложения $Ts_{\max}$ остальным предложениям аннотации,

$\max(Cs_{\max})$ — максимальное из значений в массиве $Cs_{\max}$.

Замечания

- в случае оценки «силы» смысловой связи относительно центров масс аннотаций в роли массива $Cs_{\max}$ 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 эмбединга для текста анализируемой аннотации и эмбедингов аннотаций остальных публикаций.

Утверждение 1

- результирующий рейтинг публикации, ассоциируемый с близостью её аннотации эталону, определяется произведением оценки «силы» смысловой связи работы с коллекцией и оценки смысловой связности аннотации анализируемой публикации;
- траектория навигации пользователя по коллекции строится «сверху вниз» от работы с большим рейтингом к ближайшей ей публикации с меньшим рейтингом.

Текущее состояние вопроса

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.

Очевидный недостаток данного варианта решения

Расширение аннотаций статей коллекции происходит независимо друг от друга, что критично для их взаимной оценки близости смысловому эталону.

В целях согласованности расширения аннотаций:

- переопределим оценку (1) для смысловой связности коллекции. При такой постановке $C_{S_{max}}$ содержит значения «силы» смысловой связи каждой из работ с остальными работами коллекции (относительно центров масс или полных текстов, соответственно);
- для указанной «силы» воспользуемся двумя независимыми оценками, содержащими близкими к оценке (1);
- в случае оценки относительно центров масс $C_{S_{max}}$ включает значения косинусной близости вектора центра масс анализируемой аналогичным векторам центров масс аннотаций других работ коллекции;
- при оценке относительно полных текстов $C_{S_{max}}$ состоит из значений косинусной близости эмбединга анализируемой и соответствующих векторов остальных аннотаций;
- из вариантов расширения рассмотрим два: до максимальной связности и первым из предложений объединённого *introduction* и *conclusions*, с которого было начато расширение аннотации до максимальной связности.

Замечание

Смысловая связность коллекции в целом определяется по аналогии с оценкой «силы» смысловой связи публикации с остальной коллекцией и показывает, насколько входящие в неё тексты (аннотации или рефераты, соответственно) взаимно связаны по смыслу.

Суть метода

Для каждой аннотации берётся её исходный и расширенный варианты. Далее строится декартово произведение получившихся пар вариантов аннотации для всех публикаций и выбирается элемент с наибольшим значением связности коллекции.

- 1 Каждая из n исходных аннотаций заменяется её расширенным вариантом, получается n вариантов коллекции, в каждом из которых одна из исходных аннотаций заменена расширенным вариантом.
- 2 Из получившихся n вариантов преобразования коллекции отбирается тот, который получил максимум оценки смысловой связности согласно формуле (1).
- 3 Полученное значение сравнивается со значением смысловой связности для исходной коллекции.
- 4 Если имеем рост смысловой связности, то продолжаем работу с оставшимися $(n - 1)$ аннотациями, иначе — выход.
- 5 На каждой последующей итерации в качестве исходного берётся результирующий вариант коллекции из предыдущей.
- 6 Окончательный вариант коллекции фиксируется (каждый из получившихся рефератов при этом представляется в виде полного текста и списка составляющих его предложений) с выдачей для последующей обработки.

Текущее состояние вопроса

- работа, предшествующая текущей в траектории, должна быть максимально близкой ей по смыслу из работ, имеющих больший рейтинг по сравнению с ней в анализируемой подборке;
- если максимально близкая по смыслу работа имеет меньший рейтинг, то считается, что для изучения текущей работы достаточно ознакомиться с одной из предшествующих в траектории — *возникает неоднозначность*.

Предлагаемый вариант решения проблемы

- кластеризация методом *k-means* эмбедингов полных текстов аннотаций, поиск оптимального числа кластеров — «методом локтя»;
- выбирать предшествующую публикацию в траектории из имеющих больший рейтинг среди находящихся в одном кластере с текущей;
- при этом предпочтение отдаётся работе, более близкой к центру кластера;
- если в ходе кластеризации *не получилось кластеров* с числом элементов, большим 1, то результат максимизации смысловой связности коллекции *считается неудовлетворительным*.

Замечание

Ранжируемыми текстами здесь являются результирующие версии рефератов для соответствующих вариантов коллекции с достигнутым максимумом смысловой связности.

Экспериментальный материал

Представлен статьями по разделу «Статистическая теория обучения» [сборника трудов](#) 15-й Всероссийской конференции «Математические методы распознавания образов» (11–17 сентября 2011 г., Петрозаводск).

Получение эмбедингов анализируемых текстов

Модель *SciRus-tiny* архитектуры RoBERTa, разработанная в Институте искусственного интеллекта МГУ и реализованная в системе поиска семантически схожих публикаций электронной библиотеки eLibrary.ru.

Вычисление косинусной близости эмбедингов:

- функция *cosine_similarity* библиотеки *sklearn.metrics.pairwise* — для центров масс анализируемых аннотаций;
- 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)

Исходная аннотация вместе с заголовком [Гуз И. С., ММРО-15]

Гибридные оценки полного скользящего контроля для монотонных классификаторов. Рассматривается задача обучения монотонного классификатора по выборке, которая не обязательно является монотонной. Цель работы — получение оценки полного скользящего контроля, которая могла бы быть использована для повышения обобщающей способности монотонных алгоритмических композиций.

Расширение

Цель работы состоит в получении как можно более точной верхней оценки полного скользящего контроля для монотонных классификаторов. В данной работе снимается требование монотонности выборки, и точность оценки повышается за счёт сужения семейства до множества монотонных классификаторов ближайшего соседа.

Примечание

В данном примере максимизируем связность коллекции относительно центров масс аннотаций путём расширения их исходных вариантов фразами введения и заключения (без ограничения числа максимизирующих предложений).

Таблица 1. Ранжирование исходной коллекции.

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

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

Ранжирование после расширения аннотаций

Таблица 2. Ранжирование после расширения каждой аннотации до достижения ей максимальной связности^{1,2}.

N_1	1	2	3	4	5	6	7	8	9	10
N_{13}	2	3	1	6	7	8	4	9	5	10
N_{23}	3	4	1	7	6	8	2	10	5	9
N_{14}	1	2	3	4	5	6	7	8	9	10
N_{24}	1	3	2	5	4	6	8	10	9	7

Таблица 3. Ранжирование после расширения аннотаций на одно из предложений объединённого *introduction* и *conclusions*.

N_1	1	2	3	4	5	6	7	8	9	10
N_{15}	1	3	4	5	7	8	2	10	6	9
N_{25}	1	4	2	5	7	9	3	10	6	8
N_{16}	1	2	3	4	5	6	7	8	9	10
N_{26}	1	3	2	5	4	6	8	10	9	7

Порядковые номера в ранжированных списках по близости эталону (БЭ) отвечают ранжированию аннотаций (рефератов) относительно центров масс и полных текстов, соответственно:

- N_{i3} и N_{i4} — после расширения каждой аннотации до достижения ей максимума связности;
- N_{i5} и N_{i6} — после расширения каждой аннотации первым из предложений объединённого *introduction* и *conclusions*, с которого было начато расширение до максимума связности.

i в нижнем индексе при N — вид оценки связности коллекции (центры масс или полные тексты).

¹ Светло-серый фон — результирующие аннотации (РА) совпали с исходными

² Тёмно-серый — РА совпали по разным видам расширения, но не совпадают с исходными

Кластеризация методом k -means

Таблица 4. Результаты кластеризации методом k -means (исходная коллекция).

Оптимальное количество кластеров для метода k -means		3
№ кластера	Порядковые номера (N_1) текстов, ближайших к центру кластера	
0	1, 3, 5	
1	4, 6, 2	
2	7, 10	

Таблица 5. Результаты кластеризации методом k -means после расширения каждой аннотации до достижения ей максимальной связности.

Оптимальное количество кластеров для метода k -means (оценка связности коллекции относительно центров масс аннотаций/рефератов)		2
№ кластера	Порядковые номера (N_1) текстов, ближайших к центру кластера	
0	1, 4, 3	
1	6, 10, 8	
Оптимальное количество кластеров для метода k -means (оценка связности коллекции относительно полных текстов аннотаций/рефератов)		3
№ кластера	Порядковые номера (N_1) текстов, ближайших к центру кластера	
0	1, 3, 5	
1	4, 6, 2	
2	7, 10	

Далее: $\Delta_{avg} N_{2j} (N_1)$ — средняя разница в рейтинге по БЭ для элементов кластеров (по всем, попарно, по мере удаления от центра кластера); N_{elem} — среднее число элементов в кластере.

Таблица 6. Результаты кластеризации методом k -means (расширение аннотаций на одно из предложений объединённого *introduction* и *conclusions*).

Оптимальное количество кластеров для метода k -means (оценка связности коллекции относительно центров масс аннотаций/рефератов)		4
№ кластера	Порядковые номера (N_1) текстов, ближайших к центру кластера	
0	3, 1, 5	
1	4, 7, 6	
2	10	
3	8	
Оптимальное количество кластеров для метода k -means (оценка связности коллекции относительно полных текстов аннотаций/рефератов)		3
№ кластера	Порядковые номера (N_1) текстов, ближайших к центру кластера	
0	1, 3, 5	
1	4, 6, 2	
2	7, 10	

Качественная оценка коллекции с достигнутым максимумом смысловой связности

- $\Delta_{avg} N_{2j}(N_1) / N_{elem}$, $j = \{3, 4\}$ — случаи расширения до максимальной смысловой связности без ограничения числа максимизирующих предложений;
- аналогично с $j = \{5, 6\}$ — расширение предложением объединённого *introduction* и *conclusions*, с которого было начато расширение до максимальной связности.

Для экспериментов с расширением каждой аннотации до достижения максимальной смысловой связности аннотацией при ранжировании **относительно центров масс**³ имеем:

$$\Delta_{1,4} = |3 - 7| = 4, \Delta_{4,3} = |7 - 1| = 6, \Delta_{6,10} = |8 - 9| = 1, \Delta_{10,8} = |9 - 10| = 1.$$

При этом $N_{elem} = (3 + 3) / 2 = 3$, а сама оценка

$$\Delta_{avg} N_{23} (N_1) / N_{elem} = ((4 + 6 + 1 + 1) / 4) / 3 = 1.$$

то же при ранжировании аннотаций **относительно их полных текстов**:

$$\Delta_{1,3} = |1 - 2| = 1, \Delta_{3,5} = |2 - 4| = 2, \Delta_{4,6} = |5 - 6| = 1, \Delta_{6,2} = |6 - 3| = 3,$$

$$\Delta_{7,10} = |8 - 7| = 1, \text{ при этом } N_{elem} = (3 + 3 + 2) / 3 \approx 2,67, \text{ а сама оценка}$$

$$\Delta_{avg} N_{24} (N_1) / N_{elem} = ((1 + 2 + 1 + 3 + 1) / 5) / 2,67 \approx 0,599.$$

Для экспериментов с расширением каждой аннотации на одно из предложений объединённого *introduction* и *conclusions* при ранжировании **относительно центров масс**⁴ имеем:

$$\Delta_{3,1} = |2 - 1| = 1, \Delta_{1,5} = |1 - 7| = 6, \Delta_{4,7} = |5 - 3| = 2, \Delta_{7,6} = |3 - 9| = 6.$$

При этом $N_{elem} = (3 + 3 + 1 + 1) / 4 = 2$, а сама оценка

$$\Delta_{avg} N_{25} (N_1) / N_{elem} = ((1 + 6 + 2 + 6) / 4) / 2 \approx 1,875.$$

то же при ранжировании аннотаций **относительно их полных текстов**:

$$\Delta_{1,3} = |1 - 2| = 1, \Delta_{3,5} = |2 - 4| = 2, \Delta_{4,6} = |5 - 6| = 1, \Delta_{6,2} = |6 - 3| = 3,$$

$$\Delta_{7,10} = |8 - 7| = 1, \text{ при этом } N_{elem} = (3 + 3 + 2) / 3 \approx 2,67, \text{ а сама оценка}$$

$$\Delta_{avg} N_{26} (N_1) / N_{elem} = ((1 + 2 + 1 + 3 + 1) / 5) / 2,67 \approx 0,599.$$

³ согласно данным Таблиц 2 и 5

⁴ согласно данным Таблиц 3 и 6

Основные выводы

- 1 Как следует из определения оценки, предпочтение отдаётся варианту расширения с меньшим значением отвечающего ему соотношения средней разницы в рейтинге по близости эталону для элементов кластеров и среднего числа элементов в кластере.
- 2 Поскольку кластеризация выполняется для эмбедингов полных текстов аннотаций, то решающими будут качественные оценки, полученные при ранжировании аннотаций относительно их полных текстов.
- 3 Отметим, что метод расширения на одно предложение объединённого *introduction* и *conclusions* уступает расширению до максимальной связности при ранжировании относительно центров масс.
- 4 Несмотря на вышесказанное, использование метода расширения на одно предложение объединённого *introduction* и *conclusions* на практике предпочтительнее в плане сокращения вычислительных затрат.

N_{26}	1	2	3	4	5	6	7
2							
3							
4							
5							
6							
7							
8							
9							
10							

Рис. 1. Первоначальный вариант траектории после расширения аннотаций на одно из предложений объединённого *introduction* и *conclusions*, ранжирование относительно полных текстов.

Примечание

- светло-серый фон ячеек таблицы на рисунке означает, что для ознакомления с работой, представляемой строкой, достаточно ознакомиться с одной из предшествующих работ, столбцы которых в строке выделены фоном;
- тёмно-серый фон показывает необходимость изучить предыдущую работу в траектории.

N_{26}	1	2	3	4	5	6	7
2							
3							
4							
5							
6							
7							
8							
9							
10							

Рис. 2. Для работ с порядковыми номерами $N_{26} = 3$ и $N_{26} = 6$ находим работы, лежащие с ними в одном кластере согласно данным *Таблицы 6*.

Примечание

Заметим, что для работы с порядковым номером $N_{26} = 7$ согласно *Таблице 6* нет вышестоящих по рейтингу работ из одного кластера с ней, поэтому в траектории навигации здесь остаются работы, ближайшие к центрам своих кластеров.

N_{26}	1	2	3	4	5	6	7
2							
3							
4							
5							
6							
7							
8							
9							
10							

Рис. 3. Окончательный вариант траектории навигации по коллекции.

Примечание

Поскольку работы с порядковыми номерами $N_{26} = 1$ и $N_{26} = 5$ не лежат в одном кластере с работой под номером $N_{26} = 7$, то для её изучения не требуется ознакомления с обеими вышестоящими работами, достаточно ознакомиться с одной из них. Фон соответствующих ячеек на рис. 3. оставлен светло-серым.

- ❶ Основной *результат* настоящей работы — *методика* повышения полноты изложения основного содержания научной статьи в её аннотации и заголовке с целью последующего ранжирования статей по *близости эталонному варианту* передачи смысла в их кратком изложении.
- ❷ *Принципиально новым* здесь является *предположение* о неполноте представления смысла в анализируемом тексте и возможности его воссоздания из контекста (вводного и заключительного разделов статьи) оценкой близости соответствующих многомерных векторов (эмбедингов).
- ❸ Результаты экспериментов *проиллюстрировали почти трехкратное уменьшение степени неоднозначности* при выборе предшествующей работы в траектории навигации пользователя по подборке.
- ❹ В качестве альтернативы «методу локтя» *для повышения точности* априорной *оценки* оптимального *числа кластеров* на выходе алгоритма *k-means* в рассматриваемой задаче заслуживает интерес *использование коэффициента «силуэт»* для оценки правильности отнесения образца данных к кластеру на основе среднего внутрикластерного расстояния и среднего расстояния до ближайшего кластера по каждому из образцов.
- ❺ *Представляется перспективным* здесь задействовать нейросетевые модели из ориентированных на работу с парафразами, а именно: *rut5-base-paraphraser* и *paraphrase-multilingual-MiniLM-L12-v2*.