Трансформерные модели BERT, взаимное сходство смыслов коротких текстов и их ранжирование по близости эталону

Михайлов Д. В., Емельянов Г. М.

Новгородский государственный университет
имени Ярослава Мудрого

Всероссийская конференция с международным участием
«Математические методы распознавания образов» (ММРО-2023),

12–15 декабря 2023 г.

г. Москва

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 предлагаемого решения

- «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.

Пусть

D — некоторая представительная (референтная) коллекция текстов, не являющихся сложными для выбранной аудитории читателей.

X — упорядоченная по убыванию последовательность $\mathrm{tf}\,(t,d)\cdot\mathrm{idf}\,(t,D)$ для всех слов t фразы $Ts_j \in \mathbb{T}\mathrm{s}_i$ относительно документа $d \in D$.

F — последовательность кластеров $H_1,\ldots H_r$, на которые разбивается X алгоритмом, содержательно близким алгоритмам класса FOREL.

Наибольший интерес для оценки близости фразы смысловому эталону представляют слова кластеров:

$H_1\,(X)$ — *слова-термины* исходной фразы, *наиболее уникальные* для d;

$H_{r/2}\,(X)$ — *общая лексика*, обеспечивающая *синонимические перифразы*, и *термины-синонимы*;

$H_r\,(X)$ — *слова-термины*, *преобладающие* в корпусе.

Близость отдельной фразы эталону: основные эмпирические соображения

- как можно более выраженное разделение слов на общую лексику и термины;
- слова в кластерах $H_1,\ldots H_r$, формируемых по TF-IDF слов фразы относительно некоторого $d \in D$, должны быть распределены более или менее равномерно;
- число получившихся кластеров на последовательности X должно быть как можно ближе к трём при максимуме значений TF-IDF для слов кластера H_1.

Документы в составе множества D сортируются по убыванию произведения оценок:

$$val_1 = -1/\log_{10}\left(\Sigma_{H_1}\right), \tag{2}$$

$$val_2 = 10^{-\sigma\left(\left|H_i,\, i=\{1,\, r/2,\, r\}\right|\right)}, \tag{3}$$

и, соответственно,

$$val_3 = \left|H_1 \setminus H_{r/2} \setminus H_r\right| / \mathrm{len}\left(Ts\right), \tag{4}$$

где Σ_{H_1} есть сумма величин TF-IDF слов, отнесённых к кластеру H_1 относительно $d \in D$;
$\sigma\left(\left|H_i,\, i=\{1,\, r/2,\, r\}\right|\right)$ — СКО числа элементов в кластере из списка $\{H_1,\, H_{r/2},\, H_r\}$;
$\mathrm{len}\left(Ts\right)$ — длина фразы Ts в составе группы $\mathbb{T}$s «заголовок + аннотация статьи».

Первый вариант оценки:

$$\mathrm{N}_1\left(\mathbb{T}\mathrm{s}, D\right) = \frac{\max\limits_{d\in D}\left(val_1\left(Ts_1, d\right) \cdot val_2\left(Ts_1, d\right) \cdot val_3\left(Ts_1, d\right)\right)}{\sigma\left(\max\limits_{d\in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbb{T}\mathrm{s}\right)+1}. \tag{5}$$

Здесь:

в числителе — оценка близости эталону заголовка статьи (Ts_1);
первое слагаемое в знаменателе — СКО значения близости эталону по всем $Ts_i \in \mathbb{T}$s.

Второй вариант оценки:

$$\mathrm{N}_2\left(\mathbb{T}\mathrm{s}, D\right) = \frac{\max\limits_{d\in D}\left(val_1\left(Ts_{\max}, d\right) \cdot val_2\left(Ts_{\max}, d\right) \cdot val_3\left(Ts_{\max}, d\right)\right)}{\sigma\left(\max\limits_{d\in D}\left(val_1\left(Ts_i, d\right) \cdot val_2\left(Ts_i, d\right) \cdot val_3\left(Ts_i, d\right)\right), Ts_i \in \mathbb{T}\mathrm{s}\right)+1}, \tag{6}$$

где $Ts_{\max} \in \mathbb{T}$s — фраза, по которой получен максимум близости эталону.

Утверждение 2

Максимальный итоговый рейтинг по коллекции получает статья с наибольшим значением оценки (5), попадающим в один кластер со значением оценки (6) для той же статьи.

Замечания

- элементы упорядоченной по убыванию числовой последовательности X *принадлежат одному кластеру*, если

$$\begin{cases} \left|\,\mathrm{mc}(X) - \mathrm{first}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \\ \left|\,\mathrm{mc}(X) - \mathrm{last}(X)\,\right| < \dfrac{\mathrm{mc}(X)}{4} \end{cases}, \tag{7}$$

  где $\mathrm{mc}(X)$ — центр масс последовательности как единого кластера, в качестве центра масс здесь берётся среднее арифметическое всех $x_j \in X$;

- корректное применение *Утверждения 2* предполагает отнесение к одному кластеру значений оценки (5) для статьи с максимальным итоговым рейтингом и максимального значения оценки (5) по коллекции, из которой ведётся отбор;
- в случае отсутствия в коллекции статьи, удовлетворяющей данному требованию, *максимальный итоговый рейтинг* получает статья с наибольшим значением оценки (5) по анализируемой коллекции;
- поскольку заголовок и фразы аннотации (по определению) несут некий единый смысловой образ, то допустима мена местами оценок (5) и (6) в *Утверждении 2*.

Вход: S; // последовательность текстов исходной коллекции,
// отсортированная по убыванию оценки (5)

Выход: S_{res}; // результат её ранжирования применением *Утверждения 2*

1: $S_{res} := \varnothing$;
2: **пока** $S \neq \varnothing$
3: $Flag := \text{false}$;
4: **для всех** $\mathbb{T}\text{s} \in S$
5: $Tmp := \{\text{N}_1(\text{first}(S), D), \text{N}_1(\mathbb{T}\text{s}, D), \text{N}_2(\text{first}(S), D)\}$;
6: отсортировать Tmp по убыванию;
7: **если** $\text{good}(Tmp) = \text{true}$ **то**
8: $Flag := \text{true}$;
9: $S_{res} := S_{res} \odot \{\mathbb{T}\text{s}\}$; // $\odot$ — операция конкатенации
10: $S := S \setminus \{\mathbb{T}\text{s}\}$;
11: **выход из цикла** *{для}*
12: **конец если**
13: **конец для**
14: **если** $Flag = \text{false}$ **то**
15: $S_{res} := S_{res} \odot \{\text{first}(S)\}$;
16: $S := S \setminus \{\text{first}(S)\}$;
17: **конец если**
18: **конец пока**

Здесь:

good — функция, выдающая true/false в зависимости от выполнения условия (7);

first — функция, возвращающая первый элемент заданной последовательности.

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

- *bert-base-nli-mean-tokens*;
- *sentence-transformers/distiluse-base-multilingual-cased-v1*;
- *sentence-transformers/all-MiniLM-L6-v2*;
- *sberbank-ai/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*.

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.

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

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

Исходные данные эксперимента: ранжируемые коллекции публикаций

- сборник трудов конференции «Интеллектуализация обработки информации» 2012 г., раздел «Математическая теория и методы классификации» (14 статей);
- сборник трудов конференции «Интеллектуализация обработки информации» 2010 г., раздел «Фундаментальные основы интеллектуального анализа данных» (25 статей);
- сборник трудов 14-й Всероссийской конференции «Математические методы распознавания образов» (2009 г.), раздел «Методы и модели распознавания и прогнозирования» (35 статей);
- сборник трудов 15-й Всероссийской конференции «Математические методы распознавания образов», (2011 г.), разделы «Математическая теория и методы классификации» (18 статей) и «Статистическая теория обучения» (10 статей);
- сборник трудов X Всероссийской конференции студентов, аспирантов и молодых учёных «Искусственный интеллект: философия, методология, инновации» (2017 г., 12 статей);
- коллекция по информационной безопасности и криптографии, включающая:
  - труды конференции «Распознавание-2019» (Курск, 2019 г., 1 статья);
  - труды IV Международной конференции и молодежной школы «Информационные технологии и нанотехнологии» (Самара, 2018 г., 1 статья);
  - материалы I международной молодежной научно-практической конференции «Арктические исследования: от экстенсивного освоения к комплексному развитию» (Архангельск, 2018 г., 1 статья);
  - труды международной научной конференции «Параллельные вычислительные технологии (ПаВТ'2016)» (Архангельск, 2016 г., 1 статья).

Пример: ММРО-15, Статистическая теория обучения

Таблица 1. Ранжирование статей с применением модели *bert-base-nli-mean-tokens*.

N_1	Автор (ы) и заголовок статьи	N_2	N_3
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	2	1
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	4	4
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	8	7
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	7	3
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	10	10
6	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	5	5
7	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	9	9
8	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	6	8
9	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	3	6
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	1	2

Здесь N_1 — порядковый номер статьи в ранжированном списке согласно алгоритму на *Слайде 9* относительно оценки (5); N_2 и N_3 — порядковые номера той же статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример: ММРО-15, Статистическая теория обучения

Таблица 2. Ранжирование с *sentence-transformers/distiluse-base-multilingual-cased-v1*.

N_1	Автор (ы) и заголовок статьи	N_2	N_3
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	1	1
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	2	2
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	10	10
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	3	4
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	8	7
6	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	9	9
7	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	7	8
8	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	4	5
9	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	6	3
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	5	6

Здесь N_1 — порядковый номер статьи в ранжированном списке согласно алгоритму на *Слайде 9* относительно оценки (5); N_2 и N_3 — порядковые номера той же статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример: ММРО-15, Статистическая теория обучения

Таблица 3. Ранжирование с применением модели *sentence-transformers/all-MiniLM-L6-v2*.

N_1	Автор (ы) и заголовок статьи	N_2	N_3
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	2	5
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	4	7
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	8	6
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	6	3
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	9	9
6	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	7	4
7	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	5	8
8	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	10	10
9	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	3	2
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	1	1

Здесь N_1 — порядковый номер статьи в ранжированном списке согласно алгоритму на *Слайде 9* относительно оценки (5); N_2 и N_3 — порядковые номера той же статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

Пример: ММРО-15, Статистическая теория обучения

Таблица 4. Ранжирование статей с применением модели *sberbank-ai/ruscibert*.

N_1	Автор (ы) и заголовок статьи	N_2	N_3
1	Воронцов К. В., Махина Г. А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа	4	4
2	Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов	8	8
3	Хачай М. Ю. Сходимость эмпирических случайных процессов, порождаемых процедурами обучения	10	10
4	Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированной минимизации эмпирического риска	7	6
5	Животовский Н. К. Комбинаторные оценки вероятности отклонения тестовой ошибки от ошибки скользящего контроля	5	5
6	Каневский Д. Ю. Переобучение и комбинаторная рademахеровская сложность в задачах восстановления регрессии	1	2
7	Неделько В. М. Эмпирические доверительные интервалы для условного риска в задаче классификации	9	9
8	Ботов П. В. Уменьшение вероятности переобучения итерационных методов статистического обучения	2	3
9	Ивахненко А. А., Воронцов К. В. Критерии информативности пороговых логических правил с поправкой на переобучение порогов	3	1
10	Сенько О. В., Кузнецова А. В. Системы достоверных эмпирических закономерностей в моделях оптимальных разбиений и методы их анализа	6	7

Здесь N_1 — порядковый номер статьи в ранжированном списке согласно алгоритму на *Слайде 9* относительно оценки (5); N_2 и N_3 — порядковые номера той же статьи в ранжированных списках относительно центров масс и полных текстов аннотаций, соответственно.

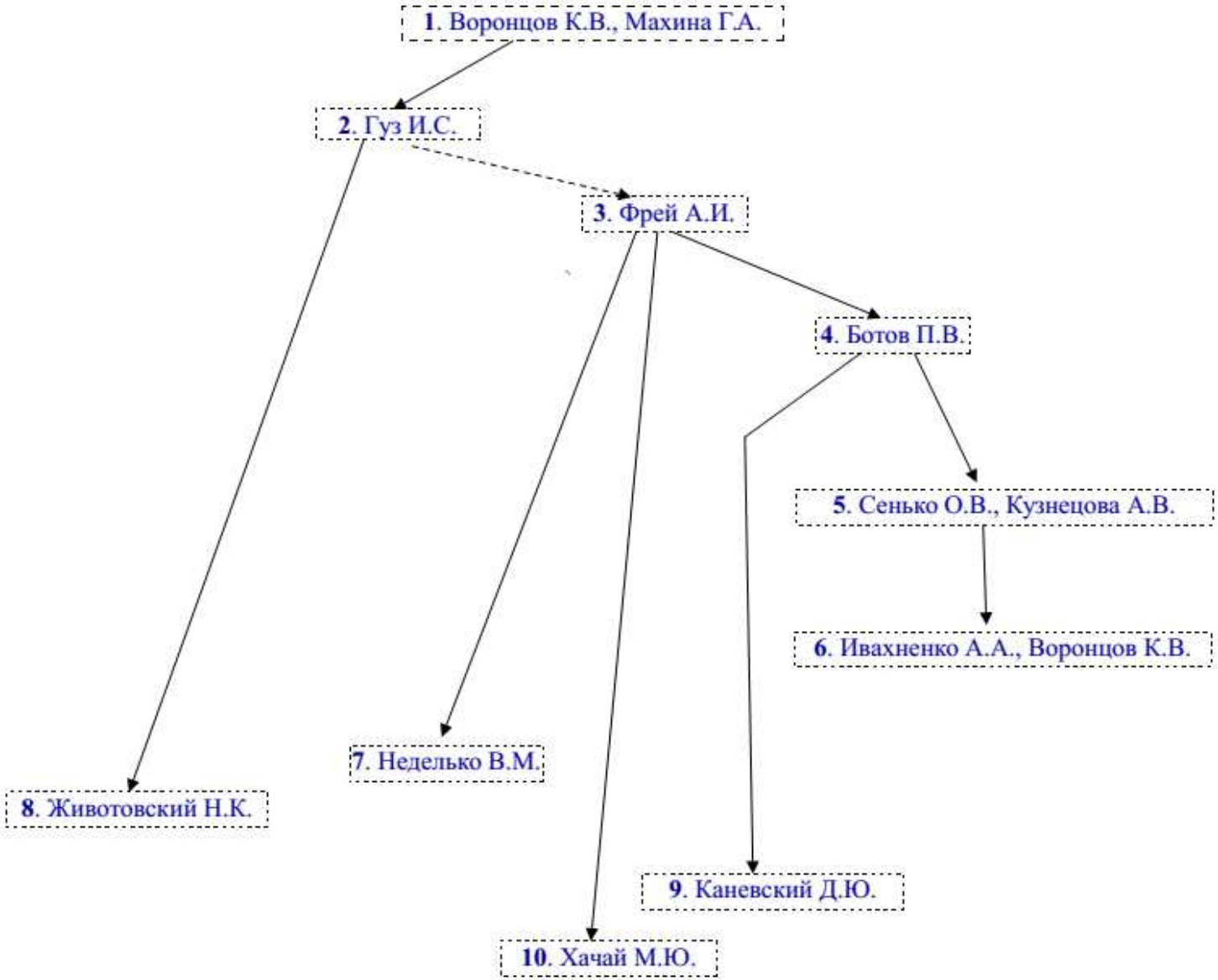

Рис. 1. Траектории навигации по коллекции (модель *sentence-transformers/distiluse-base-multilingual-cased-v1*, центры масс аннотаций).

Пример: ММРО-15, Статистическая теория обучения

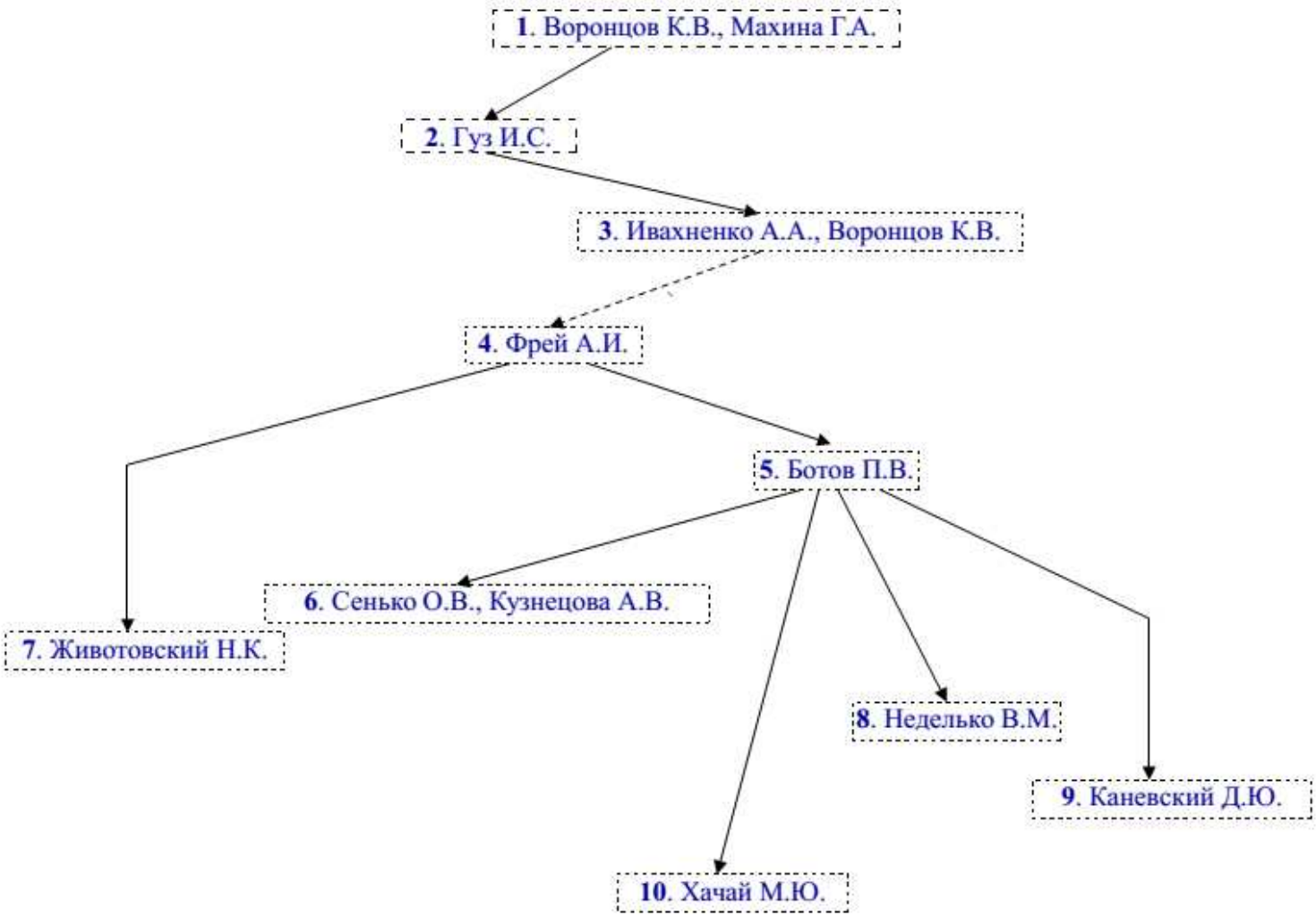

Рис. 2. Траектории навигации по коллекции (модель *sentence-transformers/distiluse-base-multilingual-cased-v1*, полные тексты аннотаций).

Пунктирная стрелка означает, что для ознакомления с работой достаточно ознакомиться с одной из предшествующих, сплошная — необходимость изучить предыдущую работу в траектории.

Основные выводы

1. В настоящей работе мы рассматриваем произвольные взаимные смысловые зависимости текстов, частным случаем которых является совпадение смыслов (семантическая эквивалентность).
2. Как и ожидалось, результаты ранжирования статей относительно разных моделей трансформеров предложений в существенной мере зависят от состава обучающей выборки, на которой обучалась модель.
3. Отметим, что максимальная близость результатам, полученным представленным на *Слайдах 6–9* методом, была характерна для экспериментов с теми моделями, которые специально обучались (или дообучались) генерировать близкие эмбеддинги для семантически близких текстов.
4. Следует отметить, что при предлагаемом ранжировании публикаций предположение об отражении аннотацией основного содержания работы и её результатов без излишних деталей может не выполняться, например, если не учитывается когнитивная сложность текста.
5. При существенном расхождении оценок по разным моделям трансформеров целесообразно расширить анализируемый материал, добавив к аннотациям вводные и заключительные разделы сравниваемых статей.
6. Отдельного рассмотрения здесь заслуживает связь вышеуказанного расхождения и близости текста смысловому эталону.

План дальнейших исследований

1. Дообучение модели ruSciBERT для задач анализа смысловой близости отдельных предложений (Sentence Similarity) и текстов (Textual Similarity) на наборах данных *https://huggingface.co/datasets/inkoziev/paraphrases* и *https://huggingface.co/datasets/merionum/ru_paraphraser*.
2. Оценка качества дообучения модели с помощью бенчмарка ruSciDocs на задачах типа рекомендации статей, похожих на статью-запрос.
3. Экспериментальные исследования дообученной версии модели ruSciBERT на задаче ранжирования научных статей по близости смысловому эталону.
4. Аналогичные эксперименты с ранжированием текстов учебной литературы из находящихся в открытом (для вузов) доступе в ЭБС «Лань».
5. Дообучение модели ruT5 для задач Sentence Similarity и Textual Similarity, сравнение с результатами по модели ruSciBERT.