Подборка публикаций по заданной теме (*плакат 2*) требует не только анализа релевантности словаря каждой публикации интересующей пользователя теме, но и учёта конечной цели самого пользователя (т.е., для решения каких именно задач делается подборка). При подготовке электронного учебного материала это означает поиск оптимального порядка работы с первоисточниками от более общего к более специфическому в целях формирования индивидуальной образовательной траектории обучаемого (студента). В идеальном случае имеем оценку 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).

В настоящей работе (*плакат 3*) задача подобного ранжирования текстов решается на основе анализа их взаимной смысловой близости с применением семейства нейросетевых языковых моделей *BERT* (от англ. *Bidirectional Encoder Representations from Transformers*). Модели данного семейства основаны на архитектуре Transformer и предварительно обучаются на больших текстовых коллекциях. С помощью указанных моделей предложения отображаются в многомерные векторы («эмбеддинги»). Содержательно каждый такой вектор показывает встречаемость заданного предложения в определённом контексте. Также возможно их построение для любого законченного текстового фрагмента, например, слова или параграфа. При этом оценка смысловой близости (т.е. «силы» смысловой связи) анализируемых текстовых фрагментов может быть формально определена через меру близости соответствующих им векторов, например, на основе косинусного расстояния. Из известных моделей семейства BERT в решаемой нами задаче наибольший интерес представляют модели типа SciBERT, обучаемые на корпусах научных текстов.

Основная идея предлагаемого решения (*плакат 4*) состоит в том, что «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. При этом среднеквадратическое отклонение оценки «силы» смысловой связи должно быть минимальным. Анализируемыми фрагментами публикаций здесь являются аннотации научных статей вместе с их 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 в настоящей работе используются две не зависящие друг от друга оценки: для полных текстов аннотаций публикаций и для центров масс аннотаций. Первым шагом по каждому предложению анализируемой аннотации для отвечающего ему эмбеддинга вычисляется массив значений косинусной близости таким же векторам остальных пред-

ложений аннотации и выбирается предложение с максимальным суммарным значением близости до остальных предложений. Такое предложение рассматривается как центр масс аннотации относительно смысловой связности.

Параллельно с оцениванием «силы» смысловой связи публикации с остальными работами в составе коллекции её аннотация проходит оценку на смысловую связность (*плакат 5*). Смысловая связность аннотации здесь предполагает то, что входящие в неё предложения должны быть максимально связаны друг с другом по смыслу. Оценка смысловой связности аннотации и оценки «силы» смысловой связи публикации с другими работами коллекции содержательно близки друг другу и имеют сходные расчётные формулы, описываемые *выражением (1)* на *плакате 5*. В случае оценки «силы» смысловой связи относительно центров масс аннотаций массив в *выражении (1)* содержательно есть 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. Результирующий рейтинг публикации, который в настоящей работе ассоциируется с близостью её аннотации эталону, определяется произведением оценки «силы» смысловой связи публикации с остальной коллекцией и оценки смысловой связности аннотации анализируемой публикации.

Для сравнения с результатами настоящей работы далее на *плакатах 6–9* приводится описание сути предложенного нами ранее метода оценивания близости текста вида «заголовок + аннотация» эталону, основанного на кластеризации слов каждой фразы анализируемого текста по значению меры TF-IDF относительно текстов представительной (референтной) коллекции, не являющихся сложными для заданной аудитории читателей. Первый из предложенных ранее вариантов оценки близости текста смысловому эталону подразумевает максимизацию близости эталону для заголовка, второй – по всем фразам. Сами документы референтной коллекции, относительно которых оценивается близость эталону, сортируются по убыванию произведения представленных на *плакате 7* оценок (2), (3) и (4), а в качестве оценки близости отдельной фразы эталону берётся наибольшее из получившихся значений. При этом (*плакат 8*) максимальный итоговый рейтинг по ранжируемой коллекции получает статья с наибольшим значением *первого варианта* оценки, попадающим в один кластер со значением *второго варианта* оценки для той же статьи.

Само ранжирование статей на основе совместного использования обоих вариантов оценки описывается алгоритмом, приведённым на *плакате 9*. Для построения иерархии документов коллекции на выходе алгоритма здесь используется аналогия с задачей вероятностного тематического моделирования, где иерархия

тем моделирует стратегию поиска с постепенным фокусированием внимания пользователя на подтемах.

Для экспериментальной апробации предложенного в настоящей работе решения были задействованы четыре представленные на *плакате 10* известные модели трансформеров предложений, работающие с русским языком, а именно:

- первая модель – *bert-base-nli-mean-tokens*;
- вторая модель – *sentence-transformers/distiluse-base-multilingual-cased-v1*;
- третья модель – *sentence-transformers/all-MiniLM-L6-v2*;
- четвёртая модель – *sberbank-ai/ruscibert*.

Программная реализация предложенного решения на Python 3.10 (включая блокнот Jupyter Notebook, исходные данные и результаты эксперимента) представлена на портале Новгородского университета. Для формирования оптимального порядка работы пользователя с публикациями уже в ранжированной коллекции для каждой из работ находится наиболее близкая ей по смыслу на основе косинусной близости соответствующих векторов-эмбеддингов.

Экспериментальный материал для апробации предложенного решения приведён на *плакате 11*. Далее на *плакатах 12–17* представлены результаты экспериментов по коллекции для раздела «Статистическая теория обучения» сборника трудов Всероссийской конференции ММРО-15 (2011 г.). Отметим, что статья из указанной коллекции, получившая максимальную близость эталону согласно алгоритму на *плакате 9* относительно заголовка, получила наибольший результирующий рейтинг также относительно первой модели (полные тексты аннотаций, *плакат 12*, зелёный фон) и второй модели (центры масс и полные тексты аннотаций, *плакат 13*, зелёный фон). Относительно первой модели (*плакат 12*, центры масс аннотаций) данная статья оказалась наиболее близкой статье с наибольшим результирующим рейтингом на основе косинусной близости соответствующих эмбеддингов (*плакат 12*, жёлтый фон). Также для статьи, максимально близкой эталону согласно алгоритму *на плакате 9*, наиболее близкая ей статья по косинусной близости эмбеддингов, получила с ней же наибольшее значение дополняемости по смыслу ранее предложенным нами методом на основе долей слов кластеров наибольших значений TF-IDF относительно:

- первой модели (*плакат 12*, полные тексты аннотаций, здесь и далее на плакатах соответствующие графы таблиц выделены оранжевым фоном);
- второй модели (*плакат 13*, центры масс и полные тексты аннотаций);
- третьей модели (*плакат 14*, центры масс и полные тексты аннотаций);
- четвёртой модели – ruSciBERT от Сбербанка (*плакат 15*, полные тексты аннотаций).

Напомним, что согласно введённому нами ранее определению дополняемость *текста 2 текстом 1* относительно их смысловых эталонов определяется долей слов кластеров наибольших значений TF-IDF по фразам *текста 1*, не входя-

щих в кластеры наибольших значений указанной меры по фразам *текста 2*, но, тем не менее, имеющих относительно тех же фраз ненулевые значения TF-IDF.

На *плакатах 16* и *17* представлены примеры траекторий навигации пользователя по коллекции, построенные (относительно центров масс и полных текстов аннотаций) согласно правилу на *плакате 10*. При этом пунктирная стрелка означает, что для ознакомления с работой достаточно ознакомиться с одной из предшествующих, сплошная – необходимость изучить предыдущую работу в траектории. Сама траектория навигации строится «сверху вниз» от публикации с большим рейтингом к публикации с меньшим рейтингом, наиболее близкой ей по смыслу.

Отметим (*плакат 18*), что в настоящей работе мы рассматриваем произвольные взаимные смысловые зависимости текстов, частным случаем которых является совпадение смыслов (семантическая эквивалентность). При этом максимальная близость результатам, полученным согласно представленному на *плакатах 6–9* методу, была характерна для экспериментов с теми моделями, которые специально обучались (или дообучались) генерировать близкие эмбеддинги для семантически близких текстов.

Следует также 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. Поэтому при существенном расхождении оценок по разным моделям трансформеров целесообразно расширить анализируемый текстовый материал, добавив к аннотациям вводные и заключительные разделы сравниваемых статей. Отдельного исследования при этом заслуживает связь указанного расхождения и близости текста смысловому эталону.

Учитывая вышесказанное, в качестве задач на дальнейшие исследования авторами выделено дообучение модели ruSciBERT для задачи анализа смысловой близости отдельных предложений (Sentence Similarity) и более общей задачи оценки близости текстов (Textual Similarity) на наборах данных перефразировок, представленных на *плакате 19*. Для оценки качества работы дообученной модели ruSciBERT здесь дополнительно будут использованы результаты решения той же задачи Textual Similarity с помощью другой известной нейросетевой модели из ориентированных на русский язык – ruT5, более точно – rut5-base-paraphraser.

Помимо научных публикаций, развиваемые в настоящей работе идеи планируется апробировать на коллекциях текстов учебной литературы из находящихся в открытом (для вузов) доступе в ЭБС «Лань». В качестве анализируемого значимого фрагмента публикации здесь планируется брать название работы + вводная часть + заключительная часть.