Подборка публикаций по заданной теме (*плакат 2*) подразумевает поиск оптимального порядка работы с первоисточниками от более общего к более специфическому. В идеале имеем оценку взаимной смысловой зависимости текстов относительно наиболее рациональных (эталонных) вариантов описания представляемых ими фрагментов знаний.

Ранее (*плакат 3*) нами было предложено решение указанной задачи на основе анализа взаимной смысловой близости текстов с применением семейства нейросетевых языковых моделей *BERT* (от англ. *Bidirectional Encoder Representations from Transformers*). Модели этого 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 рассматриваемой задаче наибольший интерес представляют модели типа SciBERT, обучаемые на корпусах научных текстов.

Основная идея (*плакат 4*): «точкой входа» в траектории работы пользователя с первоисточниками будет та публикация в составе коллекции, которая максимально связана по смыслу с остальными работами той же коллекции. При этом среднеквадратическое отклонение оценки «силы» смысловой связи должно быть минимальным. По каждому предложению анализируемой аннотации для отвечающего ему эмбеддинга вычисляется массив значений косинусной близости таким же векторам остальных предложений аннотации и выбирается предложение с максимальным суммарным значением близости до остальных предложений. Такое предложение рассматривается как центр масс аннотации относительно смысловой связности. Для самой «силы» смысловой связи публикации с другими работами коллекции используются две не зависящие друг от друга оценки: относительно полных текстов аннотаций и относительно их центров масс.

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 по остальным публикациям коллекции. При оценке «силы» смысловой связи относительно полных текстов аннотаций указанный массив будет состоять из значений косинусной близости эмбеддинга для текста анализируемой и соответствующих эмбеддингов остальных аннотаций. Результирующий рейтинг публикации, который авторами ассоциируется с близостью её аннотации эталону, определяется произведением оценки «силы» смысловой связи публикации с остальной коллекцией и оценки смысловой связности аннотации анализируемой публикации.

Но поскольку анализируемыми фрагментами публикаций в данном решении являются аннотации научных статей вместе с их заголовками как отражающие основное содержание каждой из работ и наиболее значимые результаты без излишних методологических деталей, то актуальной здесь будет проблема (*плакат 6*) полноты изложения основного содержания работы в её заголовке и аннотации. Для её решения предлагается последовательно расширять текст аннотации предложениями вводного (*introduction*) и заключительного (*conclusions*) разделов анализируемой работы согласно алгоритму на *плакате 6* с сопутствующим контролем изменения оценки смысловой связности для расширяемой аннотации.

Первым шагом вычисляется значение смысловой связности согласно формуле (1) на *плакате 5* для исходного (нерасширенного) варианта аннотации, это значение принимается за текущее. Далее в аннотацию добавляется то предложение из объединённого множества предложений введенная и заключения, для которого величина смысловой связности по расширенной аннотации будет максимальной. Если при этом новое значение смысловой связности 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. Процесс завершается тогда, когда на очередной итерации новое значение оценки (1) оказывается меньше текущего, а в качестве результата возвращается аннотация из предыдущей итерации.

Основная гипотеза здесь – степень полноты изложения основного содержания работы повышается за счёт роста смысловой связности аннотации. При этом (*плакат 7*) 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 этом берутся в исходном (нерасширенном) варианте.

При формировании оптимального порядка работы пользователя с ранжированной коллекцией 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.

В экспериментальной апробации предложенного нами решения были задействованы семь представленных на *плакате 8* известных моделей трансформеров предложений из работающих с русским языком. Помимо исходных вариантов моделей *ruscibert* и *sci-rus-tiny*, в экспериментах также участвовали варианты этих же моделей, но дообученные на датасетах перифраз *merionum/ru_paraphraser* (обе модели) и *cointegrated/ru-paraphrase-NMT-Leipzig* (только *ruscibert*). Программная реализация предложенного решения на Python 3.10 (включая блокнот Jupyter Notebook, исходные данные и результаты эксперимента) представлена на портале Новгородского университета. Экспериментальным материалом для апробации предложенного решения послужили статьи из коллекции по разделу «Статистическая теория обучения» сборника трудов Всероссийской конференции ММРО-15 (2011 г.), на которой уже иллюстрировались полученные нами ранее результаты.

На *плакатах 9–18* иллюстрируется пример расширения аннотации с номером $N_1 = 6$ по *таблице 1* на *плакате 9*. На *плакате 10* показаны полученные для неё два варианта расширения относительно разных моделей. Заметим, что второй вариант расширения нельзя назвать прямым продолжением текста исходной аннотации, хотя в данном варианте идёт речь о семействе монотонных классификаторов ближайшего соседа, но именно об этом классификаторе говорится в исходной аннотации. Но как будет далее видно из *таблицы 9* на *плакате 18*, этот вариант расширения аннотации позволил не только повысить её рейтинг относительно исходных, но и сохранить рейтинг рассматриваемой аннотации при ранжировании уже расширенных аннотаций относительно центров масс. Результаты, представленные на *плакатах 12*, *15* и *18*, иллюстрируют правильность сделанного нами предположения о том, что оценивать рейтинг расширенной аннотации необходимо именно относительно исходных вариантов остальных аннотаций ранжируемой коллекции.

Подводя итоги, следует отметить (*плакат 19*), что из 100 проведённых 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. Тем не менее, в настоящей работе остались актуальными и требуют отдельного исследования две важные проблемы. Во-первых, возможность не только расширения исходной аннотации новыми предложениями, но и удаления предложений из аннотации с заменой новыми из объединённого *introduction* и *conclusions* либо без таковой. Во-вторых, в расширенной аннотации даже при повышении рейтинга по близости эталону примерно в половине случаев новые предложения не всегда можно назвать 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.