Слайд 1

Тема настоящей работы – «Нейросетевые модели языка и отбор значимых фрагментов научных статей для безызбыточной передачи смысла».

Слайд 2

Следует отметить, что тематическая подборка публикаций предполагает поиск оптимального порядка работы с ними от более общих к более специфическим. В идеале имеем оценку взаимной смысловой зависимости текстов относительно наиболее рациональных (т. е. эталонных) вариантов описания представляемых ими фрагментов знаний.

Слайд 3

На сегодняшний день указанная задача эффективно решается с применением семейства нейросетевых языковых моделей *BERT*. Модели данного семейства основаны на архитектуре Transformer и предварительно обучаются на больших текстовых коллекциях. При этом предложения анализируемого текста отображаются в многомерные векторы, именуемые эмбеддингами. Содержательно каждый такой вектор показывает встречаемость заданного предложения в определённом контексте. Оценка близости смысла (т.е. «силы» смысловой связи) анализируемых текстовых фрагментов здесь может быть формально определена через меру близости соответствующих им векторов, например, на основе косинусного расстояния.

Слайд 4

Основная идея здесь состоит в следующем: по каждому предложению анализируемого фрагмента «аннотация + заголовок» для отвечающего ему эмбеддинга вычисляется массив значений косинусной близости аналогичным векторам остальных предложений. Далее выбирается предложение с максимальным суммарным значением близости до остальных. Такое предложение рассматривается как центр масс анализируемого фрагмента 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. При этом первой в траектории навигации пользователя по подборке будет публикация, которая максимально связана по смыслу с остальными работами в составе коллекции.

Слайд 5

Параллельно с оцениванием «силы» смысловой связи публикации с остальными работами в составе коллекции её анализируемый фрагмент проходит оценку на смысловую связность. Смысловая связность текста предполагает то, что входящие в него предложения максимально связаны друг с другом по смыслу. Оценка смысловой связности анализируемого фрагмента и оценки «силы» смысловой связи публикации с другими работами коллекции содержательно близки друг другу и имеют сходные расчётные формулы, описываемые *выражением (1)* на *слайде 5*. В случае оценки «силы» смысловой связи относительно центров масс массив в *вы-*

ражении (1) состоит из значений косинусной близости вектора центра масс фрагмента анализируемой работы аналогичным векторам по остальным публикациям в составе 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 остальных аннотаций. Результирующий рейтинг публикации, который нами ассоциируется с близостью её аннотации эталону, определяется произведением оценки «силы» смысловой связи публикации с остальной коллекцией и оценки смысловой связности анализируемого фрагмента «аннотация + заголовок. При этом траектория навигации пользователя по подборке строится «сверху вниз» от публикации с большим рейтингом к ближайшей ей по смыслу работе с меньшим рейтингом.

Слайд 6

Предложенный нами ранее вариант повышения полноты изложения основного содержания работы в её аннотации состоит в расширении текста аннотации предложениями вводного и заключительного разделов анализируемой работы с сопутствующим контролем изменения смысловой связности для расширяемой аннотации согласно последовательности шагов, представленной на данном слайде. Основная гипотеза здесь состоит в следующем: степень полноты изложения работы повышается за счёт роста смысловой связности её аннотации. Очевидный недостаток такого решения – расширение аннотаций происходит независимо друг от друга, что критично для их взаимной оценки близости смысловому эталону.

Слайд 7

Для решения указанной проблемы в настоящей работе вводится аналог оценки смысловой связности отдельного текста применительно к коллекции. При этом связность коллекции в целом оценивается по аналогии с «силой» смысловой связи публикации с остальной коллекцией и показывает, насколько входящие в неё тексты (аннотации или рефераты, соответственно) связаны друг с другом по смыслу.

Слайд 8

Максимизация смысловой связности коллекции при замене исходных аннотаций их расширенными вариантами выполняется по аналогии с максимизацией смысловой связности отдельной аннотации. Для каждой аннотации берётся её исходный и расширенный варианты. Далее строится декартово произведение получившихся пар вариантов аннотации для всех публикаций и выбирается вариант с наибольшим значением связности коллекции согласно приведённому на данном слайде алгоритму. При этом на каждой последующей итерации в качестве исходного берётся результирующий вариант коллекции из предыдущей.

Слайд 9

Рассмотрим теперь, как можно повысить точность траектории навигации по коллекции на основе результатов кластеризации эмбеддингов полных текстов аннотаций широко известным методом *k*-means. В существующем варианте решения,

работа, предшествующая текущей в траектории, должна быть максимально близкой ей по смыслу из работ с большим рейтингом в анализируемой подборке. Если максимально близкая по смыслу работа имеет меньший рейтинг, то считается, что для изучения текущей достаточно ознакомиться с одной из предшествующих работ в траектории. Для разрешения такой неоднозначности (либо как минимум снижения её степени) в настоящем исследовании предлагается выбирать предшествующую публикацию из имеющих больший рейтинг среди находящихся в одном кластере с текущей. При этом предпочтение отдаётся работе, более близкой к центру масс кластера. За основу решения нами взята представленная в открытом доступе реализация алгоритма *k*-means применительно к кластеризации эмбеддингов, использующая для поиска оптимального числа кластеров широко известный «метод локтя». Ранжируемыми текстами здесь являются результирующие версии рефератов для соответствующих вариантов коллекции с достигнутым максимумом смысловой связности. Если в ходе кластеризации не получилось кластеров с числом элементов, большим 1, то результат максимизации смысловой связности коллекции в нашем решении считается неудовлетворительным.

Слайд 10

Экспериментальным материалом для апробации предложенных решений послужили статьи по разделу «Статистическая теория обучения» сборника трудов 15-й Всероссийской конференции «Математические методы распознавания образов». Для получения эмбеддингов анализируемых текстов в настоящей работе была задействована модель SciRus-tiny известной архитектуры RoBERTa, разработанная в Институте искусственного интеллекта МГУ и реализованная в системе поиска семантически схожих публикаций научной электронной библиотеки eLibrary.ru.

Слайд 11

На данном слайде представлен пример расширения аннотации статьи из вышеупомянутого сборника до максимальной связности при максимизации смысловой связности коллекции относительно центров масс аннотаций. В данном примере аннотации расширяются предложениями вводных и заключительных разделов статей без ограничения числа максимизирующих смысловую связность предложений.

Слайд 12

На данном и последующем слайде представлены результаты ранжирования вариантов коллекции с исходными (*слайд 12*) и расширенными аннотациями (*слайд 13*). Порядковые номера в ранжированных списках по близости эталону здесь отвечают ранжированию аннотаций (либо результирующих рефератов при расширении двумя рассматриваемыми вариантами) относительно центров масс (первая цифра 1 в нижнем индексе при N), и полных текстов (первая цифра 2, соответственно).

Слайд 13

Светло-серым фоном в таблицах на *слайде 13* выделены ячейки публикаций, результирующие рефераты которых совпали с исходными аннотациями для соот-

ветствующих вариантов коллекции с достигнутым максимумом смысловой связности. Тёмно-серый фон ячеек в таблицах здесь отвечает тем публикациям, результирующие рефераты которых совпали по разным видам расширения аннотации, но не совпадают с исходными аннотациями.

Слайд 14

Таблицы 4 и *5* на данном слайде иллюстрируют результаты кластеризации методом *k*-means для эмбеддингов полных текстов, соответственно, исходных аннотаций и результирующих рефератов после расширения каждой аннотации до достижения ей максимальной связности. Результаты представлены раздельно по каждому из двух видов оценки смысловой связности коллекции – относительно центров масс и полных текстов аннотаций либо рефератов, соответственно.

Слайд 15

На данном слайде представлены результаты кластеризации при расширении аннотации первым из предложений объединённого введения и заключения, с которого начинается расширение аннотации до максимальной смысловой связности. При этом качественная оценка коллекции с достигнутым максимумом смысловой связности определяется соотношением средней разницы в рейтиинге по близости эталону для элементов кластеров из полученных методом *k*-means по эмбеддингам полных текстов аннотаций и среднего числа элементов в кластере. Более предпочтительным здесь будет вариант коллекции с меньшим значением данной величины.

Слайд 16

На данном слайде приведены результаты расчёта качественной оценки коллекции для рассматриваемого примера. Разница в рейтиинге по близости эталону у элементов кластеров при этом определяется попарно, по мере удаления от центра кластера. В рейтинге же по близости эталону предпочтение отдаётся работе, более близкой к центру масс кластера.

Слайд 17

Поскольку в предложенном нами решении кластеризация выполняется для эмбеддингов полных текстов аннотаций, то решающей будет качественная оценка коллекции, полученная при ранжировании аннотаций относительно их полных текстов. Несмотря на то, что метод расширения на одно предложение объединённого множества предложений введения и заключения уступает методу расширения до максимальной связности при ранжировании аннотаций относительно их центров масс, его использование на практике более предпочтительно в плане сокращения вычислительных затрат.

Слайд 18

На данном слайде представлен первоначальный вариант траектории после расширения аннотаций на одно предложение объединённого множества предложений введения и заключения. Светло-серый фон ячеек таблицы на рисунке означает, что для ознакомления с работой, представляемой строкой, достаточно ознакомить-

ся с одной из предшествующих работ, представляемых выделенными фоном столбцами. Тёмно-серый фон показывает необходимость изучить предыдущую работу в траектории. Напомним, «сила» смысловой связи с работой, предшествующей текущей в траектории, должна быть максимальной среди работ, имеющих больший рейтинг по близости эталону в заданной коллекции. Если максимально близкая по смыслу работа имеет меньший рейтинг, то для ознакомления с текущей работой достаточно ознакомиться с одной из предшествующих работ в траектории.

Слайд 19

На данном слайде показан промежуточный шаг по оптимизации исходной траектории навигации пользователя по подборке. Отметим, что поскольку кластеризация выполняется для эмбеддингов полных текстов аннотаций, то ранжирование аннотаций ведётся относительно их полных текстов.

Слайд 20

На данном слайде представлен окончательный вариант траектории навигации пользователя по подборке для рассматриваемого нами примера. Результаты экспериментов продемонстрировали почти троекратное уменьшение степени неоднозначности при выборе предшествующей работы в траектории навигации пользователя по заданной коллекции.

Слайд 21

Следует отметить, что используя хорошо известный «метод локтя» при поиске оптимального числа кластеров для алгоритма *k*-means, в настоящей работе мы не принимали во внимание точность оценки, вычисляемой посредством указанного метода. «Классическая» постановка алгоритма *k*-means не даёт исчерпывающего ответа на вопрос об априорной оценке оптимального числа кластеров на выходе алгоритма. В задействованном нами «методе локтя» указанная оценка вычисляется весьма приближённо ввиду отсутствия ярко выраженного пика зависимости балла кластеризации (так называемого *score*) от числа кластеров. В этом плане в качестве альтернативы «методу локтя» для повышения точности априорной оценки оптимального числа кластеров в рассматриваемой задаче заслуживает интерес использование известного 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. Также представляется песпективным задействовать для рассмотренных нами задач нейросетевые модели из реализованных для работы с парафразами с их сопутствующим необходимым дообучением для работы с научными текстами на русском языке.