Настоящая работа посвящена *взаимосвязанным проблемам* оценки близости текста наиболее рациональной (эталонной) форме передачи его смысла и формирования представительной (референтной) текстовой коллекции, относительно которой производится само оценивание.

Следует отметить, что на практике достаточно часто требуется сформировать подборку публикаций по заданной теме. Помимо научных исследований, такая задача актуальна при подготовке электронного учебного материала. При этом наибольшую значимость, как правило, имеют публикации, для которых при максимально полном раскрытии интересующей пользователя темы характерен максимум среднего числа наиболее значимых терминов в расчёте на одно простое распространённое предложение при минимуме его длины (в словах). Содержательно это соответствует максимально краткому и ёмкому изложению, отвечающему эталонному варианту передачи смысла средствами заданного языка (*плакат 2*). Данное требование можно переформулировать следующим образом: тексты в составе подборки (коллекции) должны быть максимально релевантными некоторой предметной области с точки зрения эксперта как по лексическому составу, так и по внутритекстовым связям (синтаксическим, семантическим и т.п.). Саму текстовую коллекцию при этом называют *референтной*. Для более детального анализа тексты внутри коллекции могут быть размечены по определённым правилам (синтаксически, с применением базы ролевых зависимостей и т.п.). В таком случае мы имеем дело с референтным корпусом, который может служить исходными данными для машинного обучения распознаванию зависимостей в текстах заданной тематики.

Вместе с тем, актуальной остаётся задача минимизации ручного труда эксперта при формировании подобного рода текстовых коллекций. Наиболее целесообразно здесь использование экспертом коротких текстов, которые сопоставлялись бы по лексическому составу и (возможно) по связям между словами с документами, добавляемыми в референтную коллекцию. В роли таких текстов вполне могут выступать аннотации научных статей либо другие тексты, резюмирующие значимые (с точки зрения эксперта) факты заданной предметной области. При этом имеем задачу, обратную абстрактивной суммаризации: найти текст, в котором описанные в аннотации (коллекции аннотаций) общие идеи отражены наиболее полно.

В настоящей работе представлен вариант решения данной задачи на основе долей ненулевых значений частоты слова в анализируемом документе, вычисляемых по фразам аннотаций.

В «классической» постановке (*плакат 3*) задача оценивания когнитивной сложности текста решалась на основе квантилей эмпирических распределений частоты токенов по референтному корпусу. При этом для каждого уровня языка определялся свой алфавит токенов. Так, для лексического уровня токенами являлись слова, для синтаксического – типы и длины синтаксических связей. При этом частота токена считалась аномально высокой, если она превышала 95%-й квантиль его частоты в референтном корпусе текстов, не являющихся сложными для выбранной аудитории читателей. Задача отбора документов в референтный корпус на основе коллекции аннотаций диаметрально противоположна упомянутому оцениванию когнитивной сложности текста, а именно: токены из аннотаций должны быть максимально представлены и в анализируемом документе. Для интуитивной наглядности далее в рассуждениях мы ограничимся лексическим уровнем, для других уровней языка (фонетического, морфологического, синтаксического, дискурсивного) рассуждения проводятся аналогично. В такой постановке речь идёт о минимально

необходимой представленности слов (терминов) из аннотаций в документе. Логично предположить, что в нашей задаче следует рассматривать 5%-й квантиль (т. е. 5-й перцентиль) частотной характеристики слова относительно заданного документа.

Следующий шаг – выбор самой частотной характеристики (*плакат 4*). Основное требование – независимость от числа слов документа. Будем для каждой фразы в составе каждой аннотации вычислять долю ненулевых значений TF-меры (отношения числа вхождений слова к общему числу слов документа, *term frequency*) для входящих во фразу слов относительно анализируемого документа. Одна фраза здесь соответствует простому распространённому предложению (в терминологии теории «Смысл⇔Текст»). Поскольку в реальных аннотациях доля сложных предложений минимальна, то применять данный термин к предложениям в составе аннотаций вполне допустимо. В целях максимального отражения содержания статей их аннотации будем рассматривать вместе с заголовками. При этом допускается, что одна и та же фраза может встречаться в нескольких аннотациях коллекции (например, если это статьи одного и того же автора). В любом случае каждая фраза принимается к рассмотрению только один раз.

Отметим, что использование именно доли ненулевых значений TF-меры, а не самих значений *term frequency* для слов фразы, позволяет решить проблему зависимости оценки значимости документа от числа слов в нём. Действительно, здесь важно лишь присутствие максимального количества слов из аннотаций в анализируемом документе, частота же отдельных слов здесь не принципиальна.

Базовые идеи оценивания важности документа для включения в референтную коллекцию представлены на *плакатах 5* и *6*. При этом для каждого слова каждой фразы каждой аннотации из сформированной экспертом коллекции вычисляется значение TF-меры относительно оцениваемого документа, а по отдельной фразе определяется доля ненулевых значений TF согласно формуле (2) на *плакате 5*. Далее вводится в рассмотрение 5%-й квантиль эмпирического распределения величины (2) по анализируемому документу относительно заданной коллекции аннотаций, которая при этом рассматривается как объединение множеств фраз отдельных аннотаций. Документы-кандидаты на включение в референтный корпус сортируются по убыванию значения указанного квантиля, при этом для каждого из них вводится вектор значений квантилей, куда помимо упомянутых выше 5-го и 95-го перцентилей войдут децили, а также первый и третий квартили. Для каждого из полученных векторов (*плакат 6*) вычисляется Евклидово расстояние до документа, получившего максимальное значение 5-го перцентиля эмпирического распределения доли ненулевых значений TF для заданной коллекции аннотаций. Последовательность векторов для документов-кандидатов разбивается на кластеры по величине указанного расстояния. При этом (*Утверждение 1* на *плакате 6*) наибольшую значимость для целевой коллекции будет иметь документ с максимальным значением 5-го перцентиля плюс документы кластера наименьших расстояний до него.

В целях повышения полноты (*recall*) поиска значимых документов описанную выше классификацию документов-кандидатов следует провести независимо по нескольким коллекциям аннотаций статей близких тематических направлений. Полнота поиска здесь определяется отношением числа документов, отвечающих условию *Утверждения 1* и признанных экспертом значимыми для референтного корпуса, к общему числу документов из признанных экспертом значимыми.

Точность поиска значимых документов в рассматриваемой задаче во многом зависит от состава используемой коллекции аннотаций. Содержательно здесь име-

ем требование максимизации предложенных нами ранее оценок близости эталону по каждой из аннотаций. Для оценки значимости аннотации при отборе документов в целевую коллекцию значение 5-го перцентиля эмпирического распределения, отвечающего массиву долей ненулевых значений TF, рассматривается по анализируемой аннотации и сравнивается со значением этого же перцентиля относительно объединённого множества фраз всех аннотаций коллекции (*плакат 7*). При этом среди аннотаций коллекции выделяются пять групп согласно *Утверждению 2*, из которых максимум точности поиска значимых документов дают первые три. Ранжируя аннотации внутри групп по значению указанного перцентиля, получаем альтернативный вариант оценки близости коротких текстов смысловому эталону: максимально близкими эталону будут тексты аннотаций из *группы 1*. Напомним, что предложенный нами ранее вариант оценивания близости текста эталону основан на разбиении слов каждой его фразы на классы по значению меры TF-IDF.

Экспериментальный материал для апробации предложенного подхода приведён на *плакатах 8–10*. Программная реализация на языке Python 2.7 и результаты экспериментов представлены на портале Новгородского университета. В целях более точного выделения лексического контекста слов-терминов вычисление значений *term frequency* производилось без учёта предлогов и союзов.

Далее в *таблице 1* на *плакате 11* представлены документы, отвечающие условию *Утверждения 1* и признанные экспертом значимыми по результатам экспериментов с четырьмя упомянутыми на *плакате 10* коллекциями аннотаций. Для каждого документа в таблице приведено число фраз (N_1), общее число слов с учётом всех вхождений каждого слова (N_2), число коллекций аннотаций, где документ отвечает условию *Утверждения 1* (N_3). Отметим, что среди документов-кандидатов, признанных экспертом значимыми для формирования референтного корпуса, в рассматриваемой серии экспериментов только для одного из них не выполнилось условие *Утверждения 1*, что говорит о полноте поиска, приблизительно равной 5/6. Что касается точности поиска, то в эксперименте по коллекции для раздела «Методы и модели распознавания и прогнозирования» сборника трудов конференции ММРО-14 помимо представленных в *таблице 1* документов, как отвечающий условию *Утверждения 1* был выделен упомянутый на *плакате 8* научный отчёт, не признанный экспертом значимым в решаемой задаче. По данному документу имеем $N_1 = 65$, $N_2 = 1626$, $N_3 = 1$, что в сопоставлении с остальными документами-кандидатами говорит о существенной зависимости точности поиска предлагаемым в работе методом от состава коллекции аннотаций и числа фраз в документе. Для сравнения в *таблице 2* на *плакате 11* представлены не отвечающие условию *Утверждения 1* документы-кандидаты, относительно которых предложенным нами ранее вариантом оценки удалось установить факт максимума близости эталону минимум по одной фразе в экспериментах по коллекции аннотаций для раздела «Статистическая теория обучения» сборника трудов конференции ММРО-15.

В *таблице 3* на *плакате 12* представлен результат ранжирования аннотаций в соответствии с условиями *Утверждения 2* по вышеупомянутой коллекции для раздела «Статистическая теория обучения» сборника трудов ММРО-15. Для сравнения: документ, получивший максимальное значение 5-го перцентиля эмпирического распределения доли ненулевых значений TF-меры для заданной коллекции аннотаций, здесь имеет *порядковый номер 2* по *таблице 1* на *плакате 11*. Как видно из результатов в *таблице 4* на *плакате 13*, аннотации рассматриваемой коллекции относятся к одному кластеру по величине 5-го перцентиля эмпирического

распределения доли ненулевых значений TF-меры, за исключением работы с порядковым номером 10 по *таблице 3* на *плакате 12*. Как будет видно далее в представленных на *плакатах 17* и *18* результатах экспериментов, по предложенному нами ранее варианту оценивания близости текста эталону было получено наименьшее значение указанной оценки относительно заголовка данной статьи, что служит подтверждением согласованности ранее предложенной классификации по близости эталону и классификации аннотаций согласно условиям *Утверждения 2*. Кроме того, если статьи из *таблицы 3* на *плакате 12* разбить на кластеры по величине 5-го перцентиля эмпирического распределения доли ненулевых значений TF-меры по фразам соответствующих аннотаций, добавив в разбиваемую последовательность значение указанного перцентиля по фразам аннотаций всех статей коллекции, то получим два кластера, к первому из которых будут отнесены статьи с порядковыми номерами 1 и 2, а ко второму – все остальные. В экспериментах, результаты которых представлены на *плакатах 17* и *18*, статья с номером 2 получала максимум по обеим из представленных на *плакате 14* оценок близости смысловому эталону: относительно заголовка и фразы с максимальной близостью эталону. Суть предложенного нами ранее метода оценивания близости текста эталону изложена на *плакатах 14–16*. Сами документы тематического корпуса, относительно которых оценивается близость эталону, сортируются по убыванию произведения представленных на *плакате 14* оценок (4), (5) и (6), а в качестве оценки близости отдельной фразы эталону берётся наибольшее из получившихся значений. Дополнительным подтверждением согласованности результатов в *таблицах 3–6* является пренебрежимо малая разница значений 5-го перцентиля эмпирического распределения доли ненулевых значений TF-меры по фразам аннотаций с номерами 1 и 2 из *таблицы 3*.

Основной результат настоящей работы – *метод формирования референтной текстовой коллекции для распознавания зависимостей внутри текстов определённой тематики*. При этом сами зависимости могут быть любыми и определяются целями исследования, не ограничиваясь характерной для эталонного варианта передачи смысла сочетаемостью лексических единиц и их связей. Следует отметить, что более высокую оценку значимости предлагаемым в работе методом получают документы, которые при большем числе фраз будут иметь большее среднее число наиболее значимых терминов в расчёте на одну фразу при минимуме её длины. Содержательно это соответствует более краткому и ёмкому изложению – правилу «хорошего тона» изданий по физико-математическим и техническим наукам.

Представляет интерес развитие предложенного в настоящей работе альтернативного варианта оценки близости текста смысловому эталону на основе *Утверждения 2* на *плакате 7* применительно к использованию разных документов из определяемых *Утверждением 1* на *плакате 6* по заданной коллекции аннотаций вместо документа, получившего максимум оценки значимости. Итоговая оценка близости эталону здесь будет определяться согласованием результатов классификаций аннотаций по разным документам из отобранных в корпус, например, взаимным сравнением оценок, вычисляемых в ходе проверки условий *Утверждения 2*.

В целях повышения точности поиска значимых документов заслуживает внимания адаптация предложенных оценок к другим уровням языка, помимо лексики. Сравнение классификаций по разным уровням позволит сделать вывод о значимости документа в спорных случаях, например, невыполнения условия *Утверждения 1* на одном из уровней.