Представляемая работа посвящена проблеме поиска оптимального варианта передачи смысла между экспертами и обучаемыми в системах автоматизированного обучения и контроля знаний на основе открытых тестов.

Подготовка таких тестов предполагает формирование единиц экспертных знаний, исходно представляемых текстами предметно-ограниченного естественного языка (ЕЯ). Источником знаний здесь будут публикации отечественных и зарубежных научных школ в виде монографий, обзорных статей, сборников трудов конференций и т.п. Наиболее актуальными при этом задачами являются тематическая рубрикация текстовых документов, а также представление предметных областей в виде специализированных тезаурусов и онтологий. Основная проблема – поиск варианта наиболее рациональной передачи смысла в единице знаний, определяемой множеством семантически эквивалентных (СЭ) фраз предметно-ограниченного ЕЯ. Сам смысл должен быть отражён в максимально компактном объёме текстовых данных. При этом в число задач эксперта, требующих автоматизации, входит (*плакат 3*):

- поиск СЭ-форм выражения отдельного фрагмента фактического знания в заданном ЕЯ;
- сопоставление фрагментов собственных знаний эксперта с наиболее близкими фрагментами знаний других экспертов.

Следует отметить, что решение указанных задач не сводится к простому выделению из текста понятий и отношений между ними с подсчётом семантической близости пар и групп понятий. Поиск и классификация языковых форм представления знаний здесь предполагает выявление в текстовом корпусе контекстов использования универсальной (общей) лексики, за счёт которой обеспечивается переход от исходной фразы к фразам, наиболее близким ей по смыслу. Близкую задачу, но принципиально обратного характера, решает обучаемый детектор перифраз, предложенный исследователями из Стэнфордского университета (Стэнфорд, Калифорния, США): для исходной пары фраз определяется, есть ли одна синонимичная перифраза другой. Само детектирование осуществляется нейронной сетью, для обучения которой используются результаты синтаксического разбора пар фраз из обучающей выборки, формируемой экспертом. Последняя обязательно должна содержать примеры и контр-примеры перифраз, что не вполне соответствует требованию сопоставления различных фрагментов знаний: не учитываются смысловые связи фраз помимо синонимии. Кроме того, данный подход субъективен в плане представления о самой синонимии: не учитывается предметная область каждой из фраз, а также степень их смысловой близости.

Учитывая частоту встречаемости общей лексики в текстах разной предметной ориентации, наиболее естественный путь решения вышеуказанных задач состоит в использовании известной статистической меры TF-IDF для выделения среди слов исходной фразы общей лексики и слов-терминов (в том числе в составе сочетаний). В настоящей работе рассматриваются возможности разбиения слов на классы по значению TF-IDF для поиска в текстовом корпусе описаний близких фрагментов знаний и языковых форм их выражения.

В задачах анализа текстов и информационного поиска TF-IDF есть статистическая мера, используемая для оценки важности слова в контексте документа, входящего в некоторый текстовый корпус.

Согласно классическому определению (*плакат 4*), данная мера является произведением TF-меры (отношения числа вхождений некоторого слова к общему числу слов документа) и инверсии частоты встречаемости слова в документах корпуса (IDF).

Следует отметить, что чем чаще слово встречается в документах корпуса, тем ближе к нулю будет для него значение меры IDF. Это относится как к словам общей лексики (глаголы-связки, служебные части речи), так и словам-терминам, преобладающим в корпусе. В то же время, к примеру, слова из общей лексики, задающие конверсивные замены (*«приводить ⇔ являться следствием»*) будут иметь более высокие значения меры IDF.

Допустимо предполагать (*плакат 5*), что наиболее уникальные слова в документе (с наибольшими значениями произведения мер TF и IDF) будут относиться к терминам его предметной области. Если же слово-термин имеет синонимы, встречающиеся в этом же документе, значение меры TF-меры для него будет ниже. Как и в случае вышеупомянутых конверсивных замен, здесь мы имеем меньшую встречаемость в документах корпуса каждого слова из синонимического ряда и, как следствие, более высокие значения меры IDF по сравнению со случаем отсутствия синонимов у слова.

Возьмём приведённые выше рассуждения за основу требуемого разбиения слов исходной фразы на классы по значению произведения TF и IDF.

Первый шаг (*плакат 6*) – относительно каждого документа корпуса вычислить значения TF-IDF всех слов исходной фразы. Каждая из полученных при этом последовательностей сортируется по убыванию с последующим разбиением на кластеры алгоритмом, содержательно близким алгоритмам класса FOREL. В качестве центра масс кластера здесь берётся среднее арифметическое всех его элементов. При этом оценка качества разбиения слов на классы (*формула (3) на плакате 6*) подразумевает с одной стороны как можно большее число кластеров при максимально возможном числе слов в отдельном кластере, а с другой стороны – минимум разности значения наибольшего и наименьшего диаметров кластера. Содержательно данная оценка позволяет выделить те документы текстового корпуса, относительно которых разделение слов исходной фразы на общую лексику и термины выражается в наибольшей степени.

Следующим шагом (*плакат 7*) документы корпуса сортируются по убыванию значения указанной оценки с последующим разбиением на кластеры тем же самым алгоритмом, который использовался для разделения слов исходной фразы. Здесь отбираются документы с наибольшими значениями оценки (принадлежащими первому кластеру в составе формируемой последовательности). Назовём далее эти документы лучшими по качеству. Ставится задача: из лучших по качеству документов отобрать фразы, в которых слова максимально представлены в первом, последнем и «серединном» кластерах последовательности, сформированной для исходной фразы на ос-

нове TF-IDF её слов. Введение в рассмотрение «серединного» кластера здесь необходимо (в первую очередь) для выделения общей лексики, обеспечивающей синонимические перифразы, а также терминов-синонимов. Две указанные категории лексики, не имея значения TF-IDF из первого и последнего кластеров, тем не менее, могут быть представлены в «серединном», поскольку имеют значения и TF, и IDF, близкие к средним по исходной фразе.

Как и для качества кластеризации, оценка представленности слов фразы документа в трёх указанных кластерах (*формула (4) на плакате 7*) берётся из геометрических соображений, но вместо разности значения наибольшего и наименьшего диаметров кластера здесь берётся среднеквадратическое отклонение числа слов фразы документа, представленных в кластере. Отбираемые фразы кластеризуются по значению указанной оценки, в качестве результата возвращается набор фраз, которому отвечает кластер наибольших значений.

Заметим, что предлагаемый метод не учитывает синтаксический контекст, привязка слова к нему исключила бы поиск фраз, синонимия которых исходной фразе затрагивает и синтаксис, и лексику (пример – упомянутые ранее *конверсивные замены*).

Для экспериментальной апробации предложенного метода был сформирован текстовый корпус, состав которого приведён на *плакате 8*. Тематика отбираемых работ представлена на *плакате 9*. В экспериментах по формированию единиц экспертных знаний участвовали девять исходных фраз, описывающие факты предметной области «Математические методы обучения по прецедентам» и показанные на *плакате 10*.

Программная реализация метода на языке Java и результаты экспериментов представлены на портале Новгородского университета.

В качестве примера (*плакат 11*) можно привести поиск в текстах корпуса фраз, максимально близких исходной фразе №9 по описываемому фрагменту знания и формам его выражения в русском языке. Из документов корпуса по качеству разбиения исходной фразы лучшими оказались две статьи К.В. Воронцова: в журнале «Таврический вестник информатики и математики» (№1, 2004 г.) и в сборнике трудов 15-й Всероссийской конференции «Математические методы распознавания образов». Эти два документа и послужили источником отбора фраз по максимуму оценки представленности слов в трёх наиболее значимых кластерах по TF-IDF.

Первая из результирующих фраз, представленных на *плакате 11*, служит примером связи заданного фрагмента знаний со знаниями других экспертов: определение обобщающей способности алгоритма, упоминаемой в исходной фразе, связывается здесь с понятиями вероятность ошибки и частота ошибок на контрольной выборке. Сказанное немаловажно для правильного подбора фраз, семантически эквивалентным исходным фразам №6 и 7 из представленных на *плакате 10*. Вторая фраза на *плакате 11*, является примером уже языковых выразительных средств конструирования экспертом синонимичных перифраз, ср. *«ведёт к ⇔ является результатом»*.

Следующий пример для исходной фразы №4, приведённый на *плакатах 12* и *13*, иллюстрирует поиск синонима для слова-термина *«переподгон-*

ка». Заметим, что *«переподгонка»* имеет синоним *«переобучение»* в текстах корпуса, и относительно первого из документов в таблице на *плакате 12* значение TF-IDF для него вошло в «серединный» кластер, см. для сравнения таблицы на *плакате 13*. Результирующая фраза, представленная на *плакате 12*, содержит также вариант конверсивной замены, ср. *«причина ⇔ результат»*.

Следует отметить, что если встречаемость в текстах корпуса слова-термина минимальна, то для большинства лучших по качеству кластеризации документов слово будет отнесено к последнему из кластеров по TF-IDF. Как следствие – достаточно невысокая совместная встречаемость в одной фразе с интересующими нас словами-терминами и общей лексикой, обеспечивающей синонимические перифразы. При этом фразы, близкие исходной с точки зрения эксперта по описываемому фрагменту знания либо формам его выражения, найдены не будут.

Примером может послужить слово *«заниженность»* в эксперименте по поиску фраз, максимально близких исходной фразе №8. Как видно из представленных на *плакате 14* результатов эксперимента, в нём не нашлось фраз, где помимо максимизации критерия представленности слов в трёх наиболее значимых кластерах по TF-IDF выполнялось бы требование наличия указанного слова.

Один из вариантов качественно улучшить поиск для рассматриваемого случая могло бы стать использование суммарного значения TF-IDF слов исходной фразы, встречающихся во фразе документа, в качестве альтернативы наиболее значимым кластерам по TF-IDF. Но как показал эксперимент с той же исходной фразой №8, это приводит лишь к росту числа отбираемых фраз, причём среди них не находится ни вариантов перифраз для исходной фразы, ни фраз, которые связывали бы упоминаемые в исходной фразе понятия с другими понятиями заданной предметной области.

Другой вариант – поиск слов, связанных по смыслу с заданными словами, на основе известных семантических отношений и форм их выражения в текстах. Но как видно из результатов выделения связей слов исходных фраз № 8 и 9 с помощью известной системы «Серелекс» (*плакат 15*), фиксированного набора таких знаний, как правило, недостаточно. Актуальной проблемой подобных решений является зависимость качества работы реализуемых шаблонов отношений от предметной ориентации лексики анализируемого текста. *Предложенный в настоящей работе метод* позволяет выделять понятия и отношения между ними без ориентации на предметную область и типы связей слов исходных фраз. В то же время актуальна существенная зависимость качества работы метода от подбора исходного корпуса экспертом. Одна из открытых проблем здесь (*плакат 17*) – выработка численной оценки, которая учитывала бы одновременно наиболее значимые критерии отбора документов в корпус, а именно: качество выделения тем, характер распределения терминов в теме и характер распределения тем в документе.