Представляемая работа посвящена взаимосвязанным проблемам (*плакат 2*) выделения единиц знаний из множества (корпуса) тематических текстов и отбора текстов в корпус анализом релевантности исходной фразе. Данные проблемы актуальны для построения систем обработки, анализа, оценивания и понимания информации, в частности, тестирования знаний на основе открытых тестов. Естественным источником знаний при этом будут публикации отечественных и зарубежных научных школ в виде монографий, обзорных статей, сборников трудов конференций и т.п. Конечной практической целью здесь является поиск наиболее рационального варианта передачи смысла в единице знаний, определяемой множеством семантически эквивалентных фраз предметно-ограниченного естественного языка (ЕЯ). Причём помимо отбора фраз из готового текстового корпуса, важнейшей составляющей здесь является формирование самого корпуса с включением в него публикаций, максимально релевантных рассматриваемым экспертом ситуациям действительности и языковым формам их описания. При этом (*плакат 3*):

- отбор текстов в корпус, как правило, субъективен;
- выбор критерия отбора текстов – задача нетривиальная. Здесь учитывается и уровень сложности текста, и его значимость в решаемой задаче (например, с точки зрения тематической рубрикации для составления теста по тем или иным фрагментам экспертного знания);
- значимость текста в решаемой задаче может определять выбор меры его близости исходной фразе.

Следует отметить, однако, что значимость текста здесь, как правило, безотносительна к образу, представляемому исходной фразой и выделяемому в анализируемых текстах. Сама же исходная фраза лишь в единичных случаях соответствует эталону для сопоставления. Хорошей иллюстрацией данного тезиса могут послужить случаи минимальной встречаемости слова-термина из исходной фразы в текстах корпуса. При этом по значению статистической меры TF-IDF слово не может быть безошибочно отнесено к терминам предметной области, равно как и к общей лексике, обеспечивающей переход от исходной фразы к её синонимичным перифразам. Требования к соотношениям составляющих выделяемого в тексте образа здесь можно сформулировать следующим образом (*плакат 4*):

- фрагмент анализируемого текста, отвечающий составляющей образа, отождествим с некоторой смысловой связью слов в исходной фразе;
- сила связи слов каждого такого фрагмента всегда больше силы связи любого слова данного фрагмента и слова, не принадлежащего ему;
- слабосвязанные слова исходной фразы не могут отождествляться (по определению) с одним фрагментом. Очевидно, что сочетания общей лексики и терминов исходной фразы, преобладающих в корпусе, в анализируемом тексте можно отнести к составляющим искомого образа только по присутствию фрагментов с большей силой связи слов.

Кроме того, в общем случае не выдвигается требование наличия в тексте строго заданной части составляющих образа исходной фразы (ОИФ). Поэтому корректное выделение этого образа во всех отбираемых в корпус текстах предполагает исследование встречаемости и отдельных слов, и их сочетаний с оценкой «силы» связи слов относительно текста и корпуса. В настоящей работе рассматриваются варианты такой оценки и эффективность их использования для выделения составляющих ОИФ при формировании тематического текстового корпуса.

В качестве инструментов выделения словосочетаний в задачах информационного поиска, компьютерной лексикографии, выявления плагиата, а также тематической рубрикации текстов используют (*плакат 5*) частоту *L*-грамм (по К. Шеннону), частоту и фильтрацию по тэгам, а также математическое ожидание и дисперсию. При этом само словосочетание понимается как цепочка минимум двух связанных по смыслу слов, где выбор одного слова зависит от выбора другого. Для проверки же статистической значимости словосочетания применяют методы проверки статистических гипотез, а именно: *t*-критерий Стьюдента, критерий согласия Пирсона (так называемый критерий χ^2), а также критерий отношения правдоподобия. Отметим, что данные методы для своей реализации требуют синтаксически размеченный текстовый корпус. Это необходимо (в первую очередь) для выделения биграмм, с которыми в этих методах ассоциируются словосочетания. Синтаксическая разметка текстов корпуса не поддаётся полной автоматизации и требует существенных временных затрат. Существующие же корпуса, например, Национальный корпус русского языка, в большинстве случаев не содержат требуемых данных по биграммам из анализируемых текстов.

Представленная на *плакате 6* оценка (1) для словосочетаний, выделяемых в тексте из n фраз при формировании текстового корпуса фиксированного ЕЯ, сравнима с *G*-тестом (англ. *G-test of goodness-of-fit*) для распределения Пуассона. Отметим, что корректное её применение предполагает присутствие каждого из слов пары (A,B) минимум в одной фразе анализируемого текста. Менее критичен к указанному требованию дистрибутивно-статистический метод построения тезаурусов. Этот метод содержательно наиболее близок рассматриваемой задаче выделения в анализируемых текстах образа, представляемого исходной фразой. Основная гипотеза метода заключается в наличии некоторой связи между словами, совместно встречающимися в пределах некоторого текстового интервала, в частности, в пределах одной фразы. При этом каких-либо ограничений на применяемые оценки совместной встречаемости слов не накладывается. Тем не менее, вне зависимости от вида используемая оценка, подобно (1), обычно учитывает частотные характеристики совместной встречаемости слов пары и одиночной встречаемости каждого слова.

Из оценок для «силы» связи слов в дистрибутивно-статистическом методе наиболее наглядна, но в то же время учитывает встречаемость каждого слова в отдельности, представленная на *плакате 6* оценка (2), содержательно близкая коэффициенту Танимото. Именно она и используется в настоящей работе при анализе релевантности текста исходной фразе. Для сравнения в качестве альтернативы указанной величине берётся оценка (1), где значение $sig(A,B)$ принимается равным нулю при равенстве нулю a, b, либо k.

Первый шаг (*плакат 7*) – относительно каждого документа исходного текстового множества вычисляется выбранная оценка ($sig(A,B)$ либо K_{AB}) для каждой пары слов (A,B), которым в исходной фразе соответствуют некоторые синтаксические связи. Каждая из полученных при этом последовательностей (в них включаются только ненулевые значения оценок) 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 функция ранжирования (*формула (3) на плакате 7*) подразумевает для максимально релевантных документов максимум связей с наибольшими значениями «силы» (отнесённых к кластеру «наиболее

сильных») при максимальной суммарной величине оценки «силы» для всех найденных в исходной фразе связей. Идейно оценка (3) близка предложенному в Яндекс методу определения неестественного происхождения текстового документа. Действительно, если в неестественном тексте количество редких, нехарактерных для языка сочетаний слов обычно завышено по сравнению со стандартом, а количество частых пар – занижено, то в отбираемом тексте, релевантном исходной фразе, аналогичная ситуация наблюдается с сочетаниями слов, представляемых кластером наибольших значений оценки (3).

Следующим шагом (*плакат 8*) сортировкой анализируемых документов по убыванию значений функции (3) с последующим разделением на кластеры отбираются документы с наибольшими значениями данной оценки (принадлежащими первому кластеру в составе формируемой последовательности). Назовём далее эти документы максимально релевантными исходной фразе.

Ставится задача: из максимально релевантных документов отобрать фразы, наиболее близкие исходной по представляемым фрагментам знаний. Данный шаг может быть реализован двумя способами: либо по числу найденных во фразе связей из «наиболее сильных», либо по суммарному значению «силы» указанных связей. И в том, и в другом случае отбираемые фразы кластеризуются по значению выбранной оценки, а в качестве результата возвращается набор фраз, которому отвечает кластер наибольших значений. Идейно данный подход близок исследованиям U. Manber и N. Heintze в области нахождения нечетких дубликатов текстовых документов, где в качестве меры сходства двух документов используется отношение числа общих подстрок фиксированной длины к размеру документа (в словах). Содержательно здесь имеет место разновидность контекстно-зависимого аннотирования, где одна аннотация строится сразу для нескольких документов. Следуя терминологии поисковых систем, назовём далее поиск фраз, близких исходной, в максимально релевантных ей документах построением аннотации.

Отметим, что предложенный метод оценки релевантности текста исходной фразе ограничивает рассмотрение связей слов биграммами. В случаях, когда доля общей лексики сравнима с долей терминов (например, в текстах по гуманитарным наукам), выделяемые во фразах биграммы в ряде случаев целесообразно расширять до трёх и более элементов.

Для решения указанной задачи в настоящей работе вводятся в рассмотрение *n*-граммы на последовательностях пар синтаксически связанных слов (*плакат 9*). При этом значимость *n*-граммы для ранжирования документов (*формула (4)*) оценивается из геометрических соображений и подразумевает максимизацию суммы силы связи слов в её составе при минимуме среднеквадратического отклонения указанной величины по всем связям слов в составе *n*-граммы. Значение функции ранжирования документов (*формула (5)* на *плакате 10*) здесь будет тем выше, чем большее число *n*-грамм из выделенных в исходной фразе найдено во фразах анализируемого документа при максимально возможном среднем значении суммарной силы связи слов в составе *n*-граммы с одной стороны, а с другой стороны – минимуме разности наибольшего и наименьшего из значений данной оценки для найденных *n*-грамм. Содержательно данная оценка позволяет выделить те документы исходного текстового множества, в которых составляющие образа исходной фразы в *n*-граммах представлены наиболее полно. При этом документы сортируются по убыванию значения функции ранжирования с последующим разбиением на классы тем же самым алгоритмом, который использовался для группировки найденных в

исходной фразе связей слов. Отбор фраз в аннотацию производится из документов, отвечающих кластеру наибольших значений функции ранжирования (далее – документов, лучших по *n*-граммам). Аналогично документам, но по оценке значимости для ранжирования, кластеризуются *n*-граммы относительно каждого из документов кластера наибольших значений функции ранжирования. На заключительном этапе множество фраз документов указанного кластера группируется тем же самым методом по числу слов в составе наиболее значимых *n*-грамм, а в аннотацию отбираются фразы кластера наибольших значений указанной оценки.

Экспериментальный материал для апробации предложенного метода подбирался в соответствии с критериями, представленными на *плакате 11*. Были подготовлены два варианта исходного текстового множества и, соответственно, две группы исходных фраз. Состав первого варианта представлен на *плакате 12*, исходные фразы для него – на *плакате 13*. Второй вариант приведён на *плакатах 14* и *15*, исходные фразы – на *плакате 16*.

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

На *плакате 17* приведён пример отбора фраз в аннотацию для случая, когда слова-термины из исходной фразы имеют недостаточно высокую встречаемость в анализируемых текстовых документах и, как следствие, разбиением слов исходной фразы на классы по значению меры TF-IDF удовлетворительного решения найти нельзя. Данный пример подтверждает изложенный на *плакате 4* тезис о возможности выделения в тексте сочетаний слабосвязанных слов по присутствию фрагментов с большей силой связи.

Следующий пример на *плакате 18* иллюстрирует типичные случаи не отнесения к «наиболее сильным» связей слов из разных кластеров по TF-IDF, что вполне ожидаемо, поскольку обе используемые в работе оценки силы связи слов учитывают как совместную встречаемость слов пары, так и одиночную встречаемость каждого слова. Тем не менее, как видно из таблицы на *плакате 19*, предложенный в настоящей работе метод дал меньший, чем метод на основе TF-IDF, выход фраз, не релевантных исходной ни по описываемому фрагменту знания, ни по языковым формам его выражения.

Сравнение ОИФ, выделяемого на основе оценки силы связи слов исходной фразы и на основе TF-IDF этих слов, наиболее наглядно иллюстрируется экспериментами с представленными на *плакате 13* фразами предметной области «Философия и методология инженерии знаний», где доля общей лексики больше аналогичного показателя для «Математических методов обучения по прецедентам» (*плакаты 14–16*). Как видно из таблицы в примере на *плакате 20*, большая часть общей лексики, которая могла бы обеспечивать синонимические перифразы исходной фразы, попадает в кластер наименьших значений меры TF-IDF (слова *с, который, на, в, основа, для*) относительно документа, послужившего источником отбора фраз в аннотацию. Более того, значение TF-IDF для слов указанного кластера здесь получилось равным нулю. В итоге из десяти отобранных фраз по TF-IDF слов исходной фразы не нашлось семантически эквивалентных для исходной фразы, равно как и фраз, связывающих упоминаемые в исходной фразе понятия с другими понятиями той же предметной области. Для сравнения на том же плакате представлены сочетания слов, отвечающие «наиболее сильным» связям (в порядке убывания оценки «силы»), по присутствию которых осуществлялся выбор фраз в предложенном в настоящей работе методе. Отметим, что в составе выделенных сочетаний присутствуют слова *«сложн-*

ный», *«понятие»* и *«знание»* из представляющих термины предметной области с нулевым TF-IDF. Далее на *плакате 21* для сравнения показаны фразы аннотации, значимые для соотнесения понятий *«сложная концептуальная конструкция – сложное понятие – внутренняя интерпретация»* и *«структурное описание – язык представления знаний»*, *а также «структурное описание»* и *«язык программирования высокого уровня»*. Поскольку сочетания с предлогами значимы для перифраз вида *«на основе ⇔ на базе»*, то подобные сочетания слов также не были исключены из рассмотрения. Более детально результаты экспериментов с фразами данной предметной области изложены на *плакате 22*.

В ещё большей степени точность выделения составляющих ОИФ в текстах данной предметной области повышается задействованием *n*-грамм на последовательностях пар синтаксически связанных слов. Как видно из представленного на *плакате 23* примера, при анализе биграмм по отдельности теряется существенная доля терминов, что минимизируется (как показано на *плакате 24*) в большинстве случаев именно рассмотрением *n*-грамм.

Следует отметить, что в отличие от предложенного метода, поиск фраз, близких исходной по описываемому фрагменту знания, на синтаксически размеченном текстовом корпусе, охватывающем весь заданный ЕЯ, требует предварительного выделения экспертом в исходной фразе слов и их сочетаний, представляющих термины предметной области. В качестве примера на *плакате 25* представлены слова и их сочетания для исходных фраз на *плакатах 13* и *16*, входящие минимум в одну фразу из документов Национального корпуса русского языка. Как видно из таблицы на *плакате 26*, найденные при этом фразы из понятийных связей практически не отражают синонимии. Кроме того, результативность поиска здесь зависит от представленности соответствующей тематики в текстах корпуса.

Таким образом, наряду с решением своей основной задачи, будучи совместно используемым с отбором фраз на основе TF-IDF слов исходной фразы, предложенный в настоящей работе метод позволяет автоматизировать выделение экспертом требуемых слов и их сочетаний для организации поиска в синтаксически размеченном текстовом корпусе нехудожественных текстов по заданной тематике. Кроме того, сам отбор текстов в тематический корпус на основе предложенных функций ранжирования позволяет точно задать его тему совокупностью специальных терминов предметной области, совместно встречающихся в текстовых документах. При этом в среднем в 15 раз сокращается выход фраз, не релевантных исходной фразе ни по описываемому фрагменту знания, ни по языковым формам его выражения. Здесь как перспективное направление дальнейших исследований следует отметить (*плакат 28*) выделение составляющих образа исходной фразы в текстах анализом встречаемости её слов, отвечающих кластеру наибольших значений TF-IDF, совместно с *n*-граммами на найденных синтаксических связях. В этом плане отдельного исследования заслуживает введение в рассмотрение меры TF-IDF для указанных *n*-грамм и их классификация наравне с отдельными словами. Также существенный практический интерес представляет оценка точности выделения границ ЕЯ-предложений для разных вариантов обучения классификатора.