Настоящая работа посвящена проблеме единства и целостности образа смыслового эталона, выделяемого по фразам для текста в составе тематической коллекции. Данная проблема актуальна для реализации целенаправленного отбора текстовой информации без потери полезной смысловой составляющей.

Следует отметить, что одна из ключевых задач хранения и передачи информации заключается в гарантированной поддержке корректности данных (задача обеспечения целостности). Применительно к базам знаний интеллектуальных систем это означает согласованность фрагментов знаний, независимо выделяемых из разных источников. В частности, здесь может оцениваться ситуационная близость информационных единиц, характеризуемая отношением релевантности. Содержательно именно такое отношение с оценкой его «силы» требуется выделить для независимых документов текстового корпуса, относительно которых оценивается близость смысловому эталону различных фраз одного и того же текста.

Практическим примером может послужить подготовка электронного учебного материала, где требуется систематизация экспертных знаний, представляемых текстами естественного языка. Основным требованием при этом является сортировка источников информации по степени отражения наиболее существенных понятий изучаемой предметной области при максимальной компактности и безызбыточности изложения, отвечающего эталонному варианту передачи смысла. Содержательно здесь мы имеем задачу поиска набора единиц текста и их связей, необходимого и достаточного для представления единицы знаний и отвечающего смысловому эталону. Близость текста эталону при этом необходимо оценивать без поиска перифраз. Предложенный авторами ранее подход к решению указанной задачи основан на разбиении слов каждой фразы анализируемого текста на классы по значению меры TF-IDF. При этом в роли анализируемых текстов выступали аннотации научных статей вместе с их заголовками.

Базовые идеи и предположения классификации слов исходной фразы по TF-IDF как основы оценки её близости смысловому эталону представлены на *плакатах 3–4*. Для отнесения сочетаний слов к ключевым из определяющих образ фразы в настоящей работе используется представленная на *плакате 3* интерпретация меры TF-IDF, оценивающая число одновременных вхождений всех слов анализируемого сочетания во фразы отдельного документа тематического корпуса (значение в числителе формулы (1)). При подсчете общего числа слов документа (знаменатель формулы (1)) здесь раздельно учитываются случаи совместной встречаемости слов сочетания и встречаемость без одновременного вхождения во фразу.

Используемый в работе вариант поиска необходимых и достаточных составляющих образа фразы предметно-ограниченного естественного языка в виде ключевых слов и их сочетаний, представленный на *плакате 5*, строится из следующих эмпирических соображений. Во-первых, разделение на общую лексику и термины здесь должно быть выражено как можно в большей степени, а слова в кластерах, формируемых по TF-IDF, должны быть распределены более или менее равномерно. Кроме того, число получившихся кластеров должно быть как можно ближе к трём при максимуме значений TF-IDF для слов кластера наибольших значений указанной меры. Данное требование следует понимать как максимальную релевантность терминов в составе фраз отбираемого документа сформированному корпусу. Сами документы корпуса сортируются по убыванию произведения представленных на *плакате 5* оценок, а в качестве оценки близости фразы эталону при этом берётся наибольшее из получившихся значений.

Но поскольку (*плакат 6*) для каждой фразы максимум её близости эталону достигается относительно своего документа тематического корпуса, то ключевые сочетания слов из задающих смысловые образы отдельных фраз будут выделяться относительно разных документов этого корпуса. Следовательно, указанные сочетания необходимо оценить на предмет отнесения к единому образу. В отличие от распознавания смысловых сверхфразовых единств на уровне глубинного синтаксиса, аналогичные формальные смысловые образы в рассматриваемой здесь задаче задаются неявно, а их выделение в тексте должно основываться на сопоставлении результатов классификации слов каждой его фразы по TF-IDF относительно разных документов корпуса. Как и в упомянутой задаче распознавания сверхфразовых единств, помимо терминов, выделяемые образы должны учитывать языковые выразительные средства, определяющие лучший вариант среди возможных перифраз. Основная трудность – относительно разных документов тематического корпуса одни и те же слова будут по-разному разделяться на общую лексику и термины. Помимо актуальной для вероятностного тематического моделирования проблемы разделения лексики на тематическую и общую, данный фактор в значительной мере затрудняет оценку именно выразительных средств языка.

Ставится задача оценки взаимной релевантности документов тематического корпуса, относительно которых достигался максимум близости фраз эталону, по разным фразам анализируемого текста. В настоящей работе данная задача решается путём сопоставления классификаций слов каждой фразы анализируемого текста по TF-IDF относительно разных документов заданного тематического корпуса. С этой целью для отдельной фразы вводятся в рассмотрение (*плакат 7*) векторы значений меры TF-IDF её слов относительно разных документов корпуса. Следует отметить, что TF-IDF часто используется для представления документов коллекции именно в виде числовых векторов с целью анализа их сходства, например, с помощью «мягкой» косинусной меры.

По каждой фразе анализируемого текста найденная последовательность векторов сортируется по убыванию значения расстояния до вектора того документа, относительно которого достигнут максимум близости фразы эталону. Отсортированная последовательность разбивается на кластеры по величине указанного расстояния посредством алгоритма, содержательно близкого алгоритмам класса FOREL. По факту отнесения некоторого документа корпуса к кластеру наименьших расстояний делается вывод согласно *Определению 1* на *плакате 7* о сопоставимости классификаций слов исходной фразы относительно этого документа и того документа, относительно которого достигнут максимум близости эталону, и, как следствие – о взаимной релевантности сравниваемых документов по TF-IDF. При этом (*плакат 8*) выделяемые ключевые сочетания с наибольшей вероятностью будут определять единый смысловой образ анализируемого текста, если они:

- идентифицируются как таковые относительно того документа, по которому максимум близости эталону достигался по наибольшему числу фраз в составе текстов анализируемой коллекции;
- выделяются в некоторой фразе и идентифицируются как таковые относительно некоторого документа, причём упомянутый выше документ будет относиться к кластеру наименьших расстояний до него.

Данное утверждение (*Утверждение 2* на *плакате 8*) позволяет сформулировать правило для выбора составных терминов: при прочих равных условиях из ключевых сочетаний, идентифицируемых относительно единственных документов,

отличных от упомянутого в первом пункте *Утверждения 2*, меньшее преимущество будет у сочетания, по документу которого число элементов кластера наименьших расстояний превышает половину размера корпуса. Помимо терминологических сочетаний, указанное правило позволяет выделять ключевые сочетания тематической и общей лексики, обеспечивающей перифразы. В представленных далее экспериментах примером может послужить сочетание *«учитывать эффект»*.

Введём в рассмотрение граф (*плакат 9*), в котором вершины соответствуют документам, относительно которых достигается максимум близости эталону минимум по одной фразе анализируемого текста, а каждое ребро соединяет вершины для пары взаимно релевантных по TF-IDF документов. Будем называть далее такой граф графом релевантности. Введение указанного графа позволяет сделать важный вывод по применению двух предложенных авторами ранее вариантов оценки близости текста смысловому эталону, предусматривающих минимум среднеквадратического отклонения (СКО) значения близости эталону по всем фразам анализируемого текста. Первый подразумевает максимизацию близости эталону для заголовка, второй – по всем фразам. Допустимо полагать, что указанные оценки будут точнее для того текста, граф релевантности которого – связный. Ключевые сочетания слов из задающих смысловые образы отдельных фраз также с большей вероятностью определяют единый образ текста в случае связности его графа релевантности. Данное предположение естественно согласуется с известной гипотезой о скрытых связях, согласно которой пары слов, встречающиеся в похожих моделях, стремятся иметь близкую смысловую зависимость.

Другой вывод (*Утверждение 4* на *плакате 9*) касается построения текстовой иерархии анализом встречаемости слов с наибольшими значениями TF-IDF в разных текстах коллекции (по степени дополнения эталона). Допустимо полагать, что при объединении графов релевантности всех текстов коллекции вышестоящий текст и непосредственно связанный с ним нижестоящий в формируемой иерархии должны иметь свои графы релевантности подграфами некоторой компоненты связности объединённого графа релевантности по коллекции. При прочих равных условиях при выборе вышестоящего для заданного текста предпочтение отдаётся тому, который отвечает условию данного утверждения.

В основе представленной на *плакате 9* оценки значимости документа корпуса для формирования графа релевантности текста (*формула (8)*) лежит разбиение на кластеры по расстоянию до вектора того документа, относительно которого достигнут максимум близости фразы эталону. По отдельному документу данная оценка предполагает минимум числа элементов в кластере наименьших расстояний при минимуме СКО числа элементов кластера по всем фразам анализируемого текста.

Экспериментальный материал для апробации метода приведён на *плакатах 10–12*. Программная реализация на языке Python 2.7 и результаты экспериментов представлены на портале Новгородского университета. Основным критерием при выборе коллекций, как и при подборе текстов в корпус, была максимально полная и наглядная иллюстрация разделения слов на общую лексику и термины. В целях более точного выделения смыслового контекста терминов вычисление меры TF-IDF слов анализируемых фраз производилось без учёта предлогов и союзов.

Далее в таблицах приведены результаты экспериментов по коллекции для раздела «Статистическая теория обучения» сборника трудов Всероссийской конференции ММРО-15 (2011 г.), вариант расчётной формулы для TF-IDF – «классический», использованный авторами ранее. При этом в *таблице 1* на *плакате 13* по

каждой фразе отдельного текста приводится её длина в словах и далее в скобках – номер документа по *таблице 2* на *плакате 14* для документа, относительно которого был достигнут максимум близости эталону по данной фразе. В *таблице 2* представлены те документы тематического корпуса, относительно которых максимум близости эталону был достигнут минимум по одной фразе. Для сравнения в *таблице 3* по каждому документу из *таблицы 2* приводится число фраз (по всей коллекции), а также минимальная и максимальная (в скобках) длина фразы с достигнутой относительно него максимальной близостью эталону.

В *таблице 4* на *плакате 15* представлены связи $j \to i$, у которых значения дополняемости соответствующих им текстов отличны от нуля. Непосредственно дополняемость текста $\mathbf{Ts}_j$ текстом $\mathbf{Ts}_i$ относительно их смысловых эталонов определяется долей слов кластеров наибольших значений TF-IDF фраз текста $\mathbf{Ts}_i$, не входящих в аналогичные кластеры по фразам текста $\mathbf{Ts}_j$, но, тем не менее, имеющих относительно тех же фраз ненулевые значения TF-IDF. Компоненты связности для объединённого графа релевантности по текстам $\mathbf{Ts}_j$ и $\mathbf{Ts}_i$ представлены в *таблице 4* списками вершин, а каждая вершина – соответствующим номером из *таблицы 2* на *плакате 14*. Строки для связей, где объединённый граф релевантности по текстам $\mathbf{Ts}_j$ и $\mathbf{Ts}_i$ состоит из нескольких компонент связности, выделены более тёмным фоном. Тем не менее, получившиеся компоненты являются подграфами одной и той же компоненты связности объединённого графа релевантности по коллекции. Число выполняемых для заданной связи шагов поиска наименьшей компоненты связности в объединённом графе релевантности по коллекции может служить оценкой силы связи: чем меньше шагов, тем сильнее связь.

Оценка значимости документа корпуса для формирования графа релевантности текста (*формула (8)* на *плакате 9*) может служить основой выбора вышестоящего текста для заданного в формируемой иерархии по степени дополнения смыслового эталона – как альтернатива оценке дополняемости анализом встречаемости слов с наибольшими значениями TF-IDF в разных текстах коллекции. Основная гипотеза такого выбора и подтверждающий её пример приведены на *плакате 16*.

Основной результат настоящей работы – *методика анализа* взаимной релевантности документов тематического корпуса, относительно которых оценивается близость текста смысловому эталону. Следует отметить, что говоря о взаимной релевантности документов, сравниваемых *Определением 1* на *плакате 7*, в настоящей работе мы не рассматривали возможную их мену местами с сохранением выполнимости первого из условий данного определения. В реальном тексте вероятность существования пары отличных друг от друга фраз, удовлетворяющих указанной ситуации, существенно зависит от его длины. По данной причине здесь не выдвигается возможность такой мены в качестве обязательного требования к взаимной релевантности документов. Открытая проблема – качество кластеризации документов корпуса по величине значимости для формирования графа релевантности текста. Здесь представляет интерес изучение распределения частот встречаемости документов в кластере наименьших значений *оценки (8)* на *плакате 9* по разным коллекциям одной тематики, в частности, с использованием квантилей эмпирических распределений указанных частот. Сказанное востребовано, к примеру, для определения «шумовых» документов в тематическом корпусе при использовании последнего в качестве референтного в задаче оценивания когнитивной сложности текста.