Представляемая работа посвящена проблеме передачи знаний, представляемых текстами на естественном языке (ЕЯ), между экспертами и обучаемыми в системах автоматизированного обучения и контроля знаний.

Как известно, анализ результатов открытых тестов предполагает наличие подсистемы обработки ЕЯ с учётом возможных синонимов, орфографических ошибок, отклонений от грамматической правильности предложений ответа, а также смысловой неполноты самого ответа. При этом крайне необходима двусторонняя связь «носитель ЕЯ (разработчик теста) – база знаний» с поддержкой актуального (в терминологии баз данных) состояния целостного образа отражения фрагмента действительности в сознании разработчика и в его языке. Кроме того, актуальной здесь является проблема зависимости результатов интерпретации ответа от субъективной точки зрения преподавателя-разработчика теста.

Целью исследования является разработка и теоретическое обоснование структуры знаний о синонимии, а также методов и алгоритмов их формирования и использования для совокупности задач:

- оценки схожести смыслов текстов предметно-ограниченного ЕЯ;
- автоматизации пополнения баз языковых и предметных знаний;
- поиска наиболее рационального плана передачи требуемого смысла между двумя категориями носителей заданного ЕЯ – экспертами и обучаемыми.

Основу предлагаемого решения составляет концепция *ситуации языкового употребления (СЯУ)* как единицы формализованного представления в едином контексте языковых и предметных знаний. Языковой контекст, фиксируемый указанной единицей, отражает значимые в ситуации объекты, отношения между ними и их выражения в текстах, эквивалентных по смыслу. Наиболее естественной моделью указанной единицы знаний является *формальный контекст (ФК)*, известный из теории анализа формальных понятий. При этом на основе решетки формальных понятий выделяются классы семантических отношений по сходству:

- основы синтаксически главного слова;
- флексии зависимого слова в рамках синтаксических отношений, что необходимо для их выделения и обобщения;
- лексической и флективной сочетаемости, что позволяет выявить зависимости, аналогичные смысловой связи между опорным словом и генитивной именной группой в составе генитивной конструкции русского языка.

Сами тексты, представляющие фрагменты фактического знания, объединяются в группы по сходству признаков сочетаемости слов относительно контекстов ситуаций языкового употребления. Поиск наиболее рационального плана передачи смысла между обучаемыми и экспертами при этом сводится к совокупности подзадач:

- выделение буквенных инвариантов слов (основ);
- формирование критерия информативности слов в контексте СЯУ;
- поиск множества синтаксических связей между словами и отбор максимально проективных фраз для формирования ФК эталона СЯУ.

Решение первой из задач основано на анализе частоты встречаемости букв на разных позициях относительно начала и конца слова в контексте СЯУ. Здесь реализован алгоритм выделения основ и флексий для слов в контексте СЯУ. Программная реализация алгоритма представлена на портале Новгородского университета, а также на персональной странице автора на www.machinelearning.ru. Особенность алгоритма – группировка словоформ по общности буквенного префикса, при этом его символы имеют наибольшее значение указанной частоты у словоформ группы. Одновременно выделяется общий суффикс с той же частотой встречаемости символов для случаев наличия у слова возвратных частиц.

Для решения задачи поиска наиболее компактных форм выражения заданного смысла фразами естественного языка в работе вводится *модель линейной структуры (МЛС) ЕЯ-фразы* на множестве индексов неизменных частей слов с учётом возможных синонимов (лемма 5.1). Для построения модели *смыслового эталона СЯУ* отбираются ЕЯ-фразы с максимально возможным числом наиболее информативных слов (с учётом конверсивов и синонимов) при максимальной проективности самой фразы. *Наиболее информативные слова* образуют кластер по частоте встречаемости в ЕЯ-фразах из определяющих СЯУ. Формирование синтаксических связей идёт по принципу обучения с учителем. На первом шаге по найденным паросочетаниям индексов в составе МЛС фраз, опрашивая эксперта, выделяют ложные связи относительно заданной СЯУ. По совокупности начальных знаний формируется булев вектор, который используется при проверке возможности отождествления новой связи с выделенными ранее истинными и ложными связями.

На *плакатах 12* и *13* представлен пример формирования смыслового эталона для СЯУ, описывающей связь между переобучением и эмпирическим риском.

Введение в рассмотрение смысловых эталонов для ситуаций языкового употребления позволяет минимизировать объёмы сравниваемой информации при оценке схожести смыслов ответа обучаемого и варианта ответа, сформулированного экспертом. Для сравнения на *плакате 14* представлены число объектов и признаков ситуаций языкового употребления и их эталонов по предметной области «Математические методы обучения по прецедентам». Следует отметить, что использование СЯУ в качестве единицы предварительного сжатия информации позволяет численно оценить резервируемый объём памяти для хранения текстов предметно-ограниченного ЕЯ с учётом возможных видов синонимии, что особенно актуально для программной реализации систем тестирования знаний. Традиционно за такую оценку для отдельной фразы из n слов в информатике берётся значение $vol(n) = n!$. Представленный метод выделения эталона ситуации языкового употребления позволяет оценивать данный объём сверху как $vol_1(n) = l_1 \cdot n$ и снизу как $vol_2(n) = l_2 \cdot n$, где l_1 – число семантически фраз из определяющих ситуацию языкового употребления, из которых l_2 определяют эталон (*плакат 15*).

Для вычисления оценки близости ответа обучаемого ответу, сформулированному экспертом, а также для согласования знаний, формируемых разными экспертами по заданной предметной области, вводится модель тезауруса в виде формального контекста (*плакат 16*). При этом в качестве тезаурусной единицы выступает теоретико-решёточное представление СЯУ, введённое нами ранее. Численная оценка схожести СЯУ определяется числом признаков, которые разделяются объектами сравниваемых ситуаций относительно формального контекста тезауруса. В целях минимизации объёма данных, необходимого для сохранения смысла при оценке схожести, каждая СЯУ в тезаурусе представляется смысловым эталоном. Согласование данных (*плакат 18*) об основах и флексиях слов по разным СЯУ относительно заданной предметной области, позволяет уточнить объектно-признаковые описания отдельных ситуаций и повысить точность оценок их схожести.

В качестве иллюстрации следует привести результаты работы разработанной автором демо-версии системы тестирования знаний, представленной вместе с исходным текстом на персональной странице автора на www.machinelearning.ru. На *плакате 19* представлен интерфейс системы, а также результаты выполнения открытого теста пятью испытуемыми до и после (*плакат 21*) выполнения процедуры согласования знаний. Незначительное снижение оценок близости правильному ответу на *Вопрос 4* у испытуемых *Зайцева Е.А.* и *Волкова А.В.* обусловлено заменой выделенных ранее нулевых флексий у ряда слов, представленных в тезаурусе.

В реализованной программной системе каждому заданию (вопросу) ставится в соответствие СЯУ правильного ответа (включая эталон). Для ответа, введённого испытуемым, производится поиск наиболее близкого (по буквенному составу) «правильного» варианта среди СЭ-форм. Далее идёт анализ словесных несовпадений, поиск соответствий для несовпадающих частей сравниваемых предложений в составе эталона правильного ответа и вычисление оценок близости с учётом найденных синонимов.

В целях более гибкой интерпретации ответа испытуемого численные оценки его близости правильному ответу вычисляются (*плакат 22*) для случаев неполного ответа, орфографических ошибок (из допустимых), лишних слов, которые не фигурируют в лексико-синтаксических связях, представленных в базе знаний системы.

В заключении следует отметить, что в данной работе все виды связей между главным и зависимым словом предполагались одинаково значимыми для оценки схожести фраз. Для применения таких оценок в реальных задачах оценки профессиональных знаний по отраслям определение схожести СЯУ следует переформулировать уже с позиций нечёткой логики. При этом для описания функций принадлежности потребуется системный анализ структуры профессиональных знаний в конкретной области.

Концепцию модели линейной структуры предложения также можно сделать более гибкой, введя вероятности совместной встречаемости слов относительно текстов заданной предметной области и жанра.