Формирование единиц представления предметных знаний в задаче их оценки на основе открытых тестов

Емельянов Г. М., Михайлов Д. В., Козлов А. П.

Новгородский государственный университет имени Ярослава Мудрого

Настоящая 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 (*плакат 2*) является разработка и теоретическое обоснование методов и алгоритмов поиска оптимального плана передачи смысла между экспертами и обучаемыми в системе контроля знаний с применением открытых тестов.

Основу предлагаемого решения составляет концепция *ситуации языкового употребления (СЯУ)* как единицы формализованного представления в едином контексте языковых и предметных знаний (*плакат 3*). Языковой контекст, фиксируемый указанной единицей, отражает значимые в ситуации объекты, отношения между ними и их выражения в текстах, эквивалентных по смыслу (*семантически эквивалентных, СЭ*).

Наиболее естественной моделью указанной единицы знаний является *формальный контекст (ФК)*, известный из теории анализа формальных понятий (*плакат 4*). 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.

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 (*плакат 5*) с учётом возможных синонимов (*лемма 1*). При этом (*плакат 6*) пара индексов, относительно которых задается связь, соответствует одной синтагме, а позиции индексов в МЛС позволяют определить длину связи. Ставится задача (*плакат 7*): на основе данных об абсолютных частотах встречаемости:

– отдельных индексов;
– связей (независимо от длин);
– связей, имеющих конкретную длину,

в моделях линейных структур СЭ-фраз из определяющих СЯУ найти набор минимальных текстовых единиц, необходимых и достаточных для формирования оптимального плана передачи смысла этой СЯУ.

Первый шаг — вычисление абсолютной частоты встречаемости для каждого индекса с последующей сортировкой полученной последовательности значений указанной частоты по убыванию (*плакат 8*).

Отсортированная последовательность разбивается на кластеры с применением алгоритма, содержательно близкого алгоритмам класса FOREL. Описание алгоритма на псевдокоде представлено на *плакате 9*. При этом элементы последовательности считаются принадлежащими одному кластеру, если модуль разности значений первого элемента последовательности и её центра масс меньше четверти значения центра масс, равно как и модуль разности последнего элемента последовательности и её центра масс.

В зависимости от соотношений этих абсолютных величин в очередном проходе алгоритма происходит смещение центра масс формируемого кластера либо вправо, либо влево, что достигается удалением из последовательности первого/последнего элемента с последующим его включением в выделяемый префикс/суффикс исходной последовательности. После формирования очередного кластера алгоритм рекурсивно применяется к выделенным префиксу и суффиксу, что продолжается до тех пор, пока на очередном шаге префикс и суффикс не окажутся пустыми. Программная реализация алгоритма представлена на портале Новгородского университета, `http://www.novsu.ru/file/1089439`. В качестве центра масс числовой последовательности здесь берётся среднее арифметическое значений всех её элементов. При этом (*плакат 10*) смысловой эталон СЯУ определяют те из задающих её СЭ-фраз, модели линейных структур которых включают все индексы, значения частоты встречаемости которых вошли в самый «тяжёлый» кластер (назовём их «частыми» индексами), при минимальном числе индексов, частоты встречаемости которых не вошли в этот кластер. Данное условие является *необходимым, но не достаточным* для отнесения некоторой из исходных СЭ-фраз к фразам, определяющим смысловой эталон СЯУ.

Суть следующего шага — численная оценка значимости связи слов в контексте СЯУ. Предлагаемый метод оценивания идейно близок методу опреде-

ления неестественного происхождения текста, основанному на изучении статистики встречаемости пар соседних слов в тексте и реализованному в Яндекс (см. Гречников Е. А., Гусев Г. Г., Кустарев А. А., Райгородский А. М.).

Для оценки «силы» связи слов (вне зависимости от их взаимного расположения в линейном ряду фразы) вводится весовая функция, представленная на *плакате 11*. При этом формируется упорядоченная по убыванию последовательность значений указанной функции для индексных пар, выделенных на множестве моделей линейных структур СЭ-фраз из определяющих СЯУ, преобразованном согласно *лемме 1*. Сформированная последовательность разбивается на кластеры описанным выше методом с применением алгоритма, представленного на *плакате 9*. При этом связи, максимально значимые для формирования искомых единиц знаний, будут иметь значения весовой функции, вошедшие в самый «тяжёлый» кластер. По аналогии с «частыми» индексами назовём далее такие связи «весомыми».

Связи, не попавшие в категорию «весомых», группируются вышеописанным методом по величине среднеквадратического отклонения длины связи (СКОДС) относительно рассматриваемого множества моделей линейных структур (*плакат 12*). На данном этапе формируется множество кандидатов на роль вершины синтаксического дерева для каждой из исходных СЭ-фраз. В качестве основополагающей здесь выдвинута следующая гипотеза: индекс, соответствующий вершине, должен входить в одну из связей кластера наименьших значений СКОДС и одновременно в связь, относящуюся к некоторому другому кластеру из полученных по указанной величине. При этом индекс, отвечающий вершине, не входит ни в одну из «весомых» связей.

Введением группировки по СКОДС *достаточное условие* для отнесения фразы к определяющим смысловой эталон СЯУ может быть сформулировано следующим образом (*плакат 13*): помимо выполнения *необходимого* условия, представленного на *плакате 10*, а также минимальной длины флективной части (флексии) каждого слова в составе фразы, МЛС фразы должна иметь минимум индексов, не вошедших в группу «частых», не фигурирующих в составе «весомых» связей и не являющихся кандидатами на роль вершины синтаксического дерева ни для одной из исходных СЭ-фраз.

Последовательным отбором фраз, отвечающих *необходимому* и *достаточному* условиям отнесения к определяющим эталон, решается задача выбора максимально компактного объёма текстовых данных из исходного множества СЭ-фраз для передачи смысла, соответствующего СЯУ. Заметим, что описанная методика не учитывает проективности ЕЯ-фраз, поскольку последняя сама по себе не гарантирует сохранение синтаксических групп, что исключает введение искусственного ограничения на проективность при выделении синтагматических связей согласно предложенному в работе принципу.

Предложенный метод формирования смыслового эталона был апробирован на материале ЕЯ-описаний шести фактов предметной области «Математические методы обучения по прецедентам». Программная реализация метода на языке Visual Prolog 5.2 вместе с исходными кодами и результатами экспериментов представлена на портале Новгородского государственного университета имени Ярослава Мудрого, `http://www.novsu.ru/file/1089439`.

Для выделения основ и флексий слов, составляющих исходные СЭ-фразы, была реализована группировка словоформ в рамках СЯУ по общности префикса и (при необходимости) суффикса.

Основная идея предложенного метода выделения основ и флексий в составе слов состоит в том, что символы общего префикса у различных форм одного и того же слова в контексте СЯУ имеют максимальную встречаемость для своих позиций в слове. Такая же частота встречаемости будет и у символов общего суффикса, соответствующего, в частности, возвратным частицам. При этом суммарная длина общих префикса и суффикса должна составлять минимум треть длины слова, а разность длин любой пары слов, имеющих общий префикс (как в совокупности с общим суффиксом, так и без него), всегда меньше половины длины меньшего слова. Ключевые процедуры и функции алгоритма, реализующего данный метод, приведены на *плакате 14*, сам алгоритм представлен на *плакате 15*.

Исходные данные экспериментов приведены на *плакате 16*. В них число СЭ-фраз, задающих СЯУ, варьировалось в диапазоне от 6 до 56, а число слов во фразе — от 5 до 18.

Следует отметить (*плакат 17*), что для каждой найденной связи слов её направление в текущей реализации задаётся экспертом. При этом направление может быть задано только для тех связей, которые будут определены экспертом как истинные. Сформированным таким образом знаниям системы об истинных и ложных связях в рамках отдельной СЯУ соответствует булев вектор, где часть компонент отождествляется с истинными, а другая часть — с ложными связями. Совокупность знаний в виде указанных векторов по разным СЯУ может использоваться для изучения закономерностей совместной встречаемости слов в составе лексико-синтаксических связей, описываемых, в частности, формальным контекстом на *плакате 4*.

Пример (фрагмент) исходного множества СЭ-фраз, задающих СЯУ № 1 из представленных на *плакате 16*, и смысловой эталон указанной СЯУ показаны на *плакатах 17–21*. Для сравнения в таблицах на *плакате 22* по найденным связям в рамках эталона СЯУ представлены значения весовой функции и среднеквадратического отклонения длины связи относительно множества моделей линейных структур исходных СЭ-фраз, преобразованном согласно *лемме 1*. Данные о кластерах, выделенных по значению среднеквадратического отклонения длины связи, приведены на *плакате 23*.

Следует отметить, что введённая концепция смыслового эталона при использовании СЯУ в качестве единицы предварительного сжатия информации позволяет оценить резервируемый объём памяти для задач текстовой обработки с учётом возможных видов синонимии. Традиционно за такую оценку для отдельной фразы максимальной длиной в n слов берётся значение $\mathrm{vol}(n) = n!$. Предложенный метод формирования эталона СЯУ позволяет дать оценку данного объёма сверху как $\mathrm{vol}_1 = n_1 l_1$ и снизу как $\mathrm{vol}_2 = n_2 l_2$, где l_1 — число СЭ-фраз из задающих СЯУ, из которых l_2 определяют эталон, n_1 и n_2 — максимальная длина фразы по СЯУ в целом и из определяющих эталон, соответственно. Соотношение указанных оценок для СЯУ из представленных на *плакате 16* приведено в таблице на *плакате 24*.

В заключении отметим, что предложенный метод выделения смысловых эталонов даёт *минимум четырёхкратное* сокращение объёма текстовых данных, необходимых для передачи единицы знаний посредством ЕЯ без потери полезной составляющей между экспертами и обучаемыми в открытых тестах.

Наиболее *слабым местом* предложенного решения является относительно малый объём исходных данных для вычисления исследуемой характеристики связи слов — среднеквадратического отклонения её длины. Здесь как перспективное направление дальнейших изысканий следует отметить согласование данных об основах и флексиях, выделяемых по разным СЯУ относительно фиксированной предметной области. Сказанное позволит дополнительно в среднем на 1,5% сократить объём баз знаний, формируемых на основе предложенного авторами метода.

В настоящей работе допустимость выделяемых связей слов, а также их направление задаются экспертом вручную. Следует отметить, что с учётом особенностей ЕЯ-форм ответов на тестовые задания такие трудозатраты являются вполне оправданными. Привлечение внешних морфологических и синтаксических анализаторов для рассматриваемого круга практических задач потребовало бы существенно больших трудозатрат, в частности, по изучению результатов разбора и их коррекции с учётом особенностей того или иного предметно-ограниченного ЕЯ-подмножества. Здесь отдельного внимания заслуживает использование статистических характеристик признаков словоформ для определения главных и зависимых слов в составе связей относительно СЯУ в предметно-ограниченном подмножестве естественного языка. В частности, существенный практический интерес представляет интерпретация широко известной меры TF-IDF для оценки важности слова в контексте СЯУ. При этом воссоздание целостного образа отдельной ситуации употребления предметно-ограниченного ЕЯ-подмножества наиболее целесообразно вести на основе вероятностей совместной встречаемости лексико-синтаксических связей слов в текстах этого языкового подмножества. Сама совокупность СЯУ по заданной предметной области здесь будет выступать в роли коллекции текстовых документов.