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 поиска необходимого и достаточного набора признаков единицы знаний, представляемых текстами естественного языка (ЕЯ) и оцениваемых с применением теста открытой формы.

Как известно, эффективная реализация открытых тестов предполагает известную структуру ЕЯ-форм выражения знаний эксперта. Выделение таких форм требует изучения семантически эквивалентных (СЭ) описаний одного и того же факта заданной предметной области на конкретном естественном языке. Причём сама интерпретация результата теста отнюдь не сводится к простому обнаружению парафраз «ответ испытуемого–правильный ответ». Основная *проблема* здесь — поиск наиболее рационального плана передачи смысла экспертом в «правильном» ответе, сам же смысл в итоге должен быть отражён в максимально компактном объёме текстовых данных. Именно относительно этих данных и оценивается правильность ответа испытуемого.

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.

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 (*плакат 5*):

- выделение неизменяемых частей (основ) и изменяемых частей (флексий) для слов в составе исходных СЭ-фраз;

– формирование критерия информативности слов в контексте СЯУ;
– выделение и классификация связей слов в составе фраз, задающих 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 6*) с учётом возможных синонимов (по *определению 3*).

Первый шаг — вычисление абсолютной частоты встречаемости для каждого индекса с последующей сортировкой полученной последовательности значений указанной частоты по убыванию (*плакат 7*).

Отсортированная последовательность разбивается на кластеры с применением алгоритма, содержательно близкого алгоритмам класса FOREL. Описание алгоритма на псевдокоде представлено на *плакате 8*. 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 9*) 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. Действительно, *Утверждение 1* затрагивает исключительно лексический состав отбираемых фраз, не принимая во внимание связи слов. Как следствие, при отборе фраз не учитывается синонимия, затрагивающая одновременно и синтаксические связи, и лексику (ср. *плакат 10*).

Суть следующего шага — численная оценка значимости связи слов в контексте СЯУ. Предлагаемый метод оценивания идейно близок методу определения 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.). При этом (*плакат 11*) вводятся в рассмотрение абсолютные частоты встречаемости связей: независимо от длин и имеющих конкретную длину в моделях линейных структур СЭ-фраз из определяющих СЯУ.

Для оценки «силы» связи слов (вне зависимости от их взаимного расположения в линейном ряду фразы) вводится весовая функция, представленная на *плакате 12*. 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 *определению 3*. Сформированная последовательность разбивается на кластеры описанным выше методом с применением алгоритма, представленного на *плакате 8*. 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».

Заметим (*плакат 13*), что не предполагая никаких изначальных гипотез относительно смысловой связи слов, данная оценка тем не менее зависит от частоты встречаемости каждого из них в анализируемых СЭ-фразах. А это значит, что «весомыми» будут связи только между словами, рассматриваемыми *Утверждением 1* в качестве основы отбора «эталонных» фраз.

Данная проблема могла бы быть решена по аналогии с выделением лексико-синтаксических шаблонов на основе n-грамм, но при условии априори известных возможных значений n. С учётом отсутствия таких данных было использовано следующее предположение: из тех связей, которые не вошли в число «весомых», наименьший разброс длины будет у связей, затрагивающих вершины синтаксических деревьев анализируемых фраз. Отметим, однако, что с учётом возможности свободного порядка слов во фразе обратное утверждение будет не всегда верным.

Связи, не попавшие в категорию «весомых», группируются описанным выше методом по величине среднеквадратического отклонения длины связи (СКОДС) относительно рассматриваемого множества моделей линейных структур (*плакат 14*). В качестве основополагающей здесь выдвинута следующая гипотеза: индекс, отвечающий 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 вершины» 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 15*): помимо выполнения *необходимого* условия, представленного на *плакате 9*, 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 связей слов 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`.

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.

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 17*, сам алгоритм представлен на *плакате 18*.

Следует отметить (*плакат 19*), что направление каждой найденной связи слов в текущей реализации задаётся экспертом. При этом направление может быть задано только для тех связей, которые будут определены экспертом как истинные. 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. Заметим, что формируемые знания касаются связей исключительно в рамках эталона СЯУ. Ограничение выделяемых связей указанными рамками обусловлено использованием именно этих связей для оценки близости ответа испытуемого правильному ответу.

Пример (фрагмент) исходного множества СЭ-фраз, задающих СЯУ № 1 из представленных на *плакате 16*, и смысловой эталон указанной СЯУ показаны на *плакатах 20–22* и *27*. Фраза минимальной длины из определяющих СЯУ выделена зелёным цветом на *плакате 20*, фраза максимальной длины — красным на *плакате 21*. Для сравнения в таблицах на *плакате 23* 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 *определению 3*. Данные о кластерах, выделенных по значению среднеквадратического отклонения длины связи, приведены на *плакате 24*.

При этом к основам слов-кандидатов на роль вершин синтаксических деревьев были отнесены: *«привод/ведет»*, *«связан»*, *«результат/следстви»*, *«причин»*, *«котор»*, *«есть/явля/служ»*, а также предлоги *«с»* и *«к»*. Как видно из таблицы на *плакате 25*, связи, вошедшие в кластер наименьших значений СКОДС и затрагивающие потенциальные вершины синтаксических деревьев, соответствуют тем фрагментам анализируемых фраз, которые минимально изменяются при перифразировании. Примеры таких фрагментов приведены на *плакате 26* для СЯУ №1 из представленных на *плакате 16*.

Выделение потенциальных вершин деревьев синтаксического подчинения фраз, определяющих эталон, позволяет минимизировать информацию, запрашиваемую у эксперта, при определении направлений для связей слов в контексте отдельной СЯУ. При этом число кластеров, выделенных по значению СКОДС и используемых для определения кандидатов на роль вершин синтаксических деревьев, должно быть минимум два. С другой стороны, число исходных СЭ-фраз, определяющих СЯУ, является конечной величиной и в первую очередь зависит от числа возможных синонимов на лексическом и синтаксическом уровне в рассматриваемом ЕЯ-подмножестве. Типы же смысловых связей слов в рамках отдельной фразы изначально не оговариватся, поэтому для полноты учёта смыслового контекста СЯУ, ограниченного набора известных семантических отношений и форм их выражения в текстах, как правило, недостаточно. Подтверждение тому — резльтаты поиска смысловых связей с помощью системы «Серелекс» (`http://serelex.cental.be`) между словами, выделенными согласно *Утверждениям 1* и *2*, в качестве основы формирования эталона СЯУ №1 из представленных на *плакате 16*.

Заметим, что из найденных отношений только три (*«risk — result»*, *«risk — reason»* и *«риск — с»*) связывают слова из представленных в таблице на *плакате 29*, причём ни одна из указанных связей не имеет синтаксической природы в контексте рассматриваемой СЯУ, что необходимо для формирования объектно-признаковых связей в составе модели эталона СЯУ. Более того, связь *«риск — с»* по указанной причине в примере на *плакате 27* определена экспертом как нерелевантная (ложная), как и связь *«risk — reason»* (*«риск — причина»*). Таким образом, предложенный в настоящей работе метод формирования смыслового эталона СЯУ, не будучи ориентированным на определённые типы связей между словами исходных СЭ-фраз, позволяет их выделять более точно и на основе меньших обучающих выборок применительно к задаче формирования единиц знаний для открытых тестов.

Следует отметить, что введённая концепция смыслового эталона СЯУ позволяет оценить объём текстовой информации, необходимой для передачи единицы знаний посредством ЕЯ без потери полезной составляющей и с учётом возможных видов синонимии. Предложенный метод формирования эталона СЯУ даёт такую оценку 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 30*.

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.

Экспериментальным подтверждением соответствия рациональной передаче смысла является безошибочность разбора «эталонной» фразы парсером, ориентированным на наиболее вероятные в языке модели словосочетаний и предложений при единственности компоненты связности графа разбора. С учётом того, что при отборе «эталонных» фраз рассматриваются все возможные порядки слов, выделяемых согласно *Утверждениям 1* и *2*, здесь в дальнейшем может потребоваться классификация отбираемых фраз по значению суммарной длины связей слов в их составе.

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 реконструкцию целостного образа СЯУ в виде совокупности определяющих её СЭ-фраз и смыслового эталона на основе текстов тематического корпуса. Основой такого решения может стать численная оценка возможности совместного появления лексико-синтаксических связей во фразе с использованием оценочной функции, аналогичной весовой функции для отдельной связи.

Другая немаловажная задача здесь — 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 предложенным в работе методом.

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 направлений связей слов в ситуациях употребления предметно-ограниченного подмножества естественного языка. В частности, существенный практический интерес представляет интерпретация известной меры TF-IDF для оценки важности слова в контексте СЯУ. Сама совокупность СЯУ по заданной предметной области здесь будет выступать в роли коллекции текстовых документов.