

Формирование и кластеризация контекстов для существительных русского языка в рамках Расщепленных Значений.

Д. В. Михайлов, Г. М. Емельянов, Н. А. Степанова

Новгородский Государственный Университет имени Ярослава Мудрого

Настоящая работа посвящается (*Плакат 1*) решению *проблемы* выявления Расщепленных Значений (РЗ) в процессе семантической кластеризации текстов Естественного Языка (ЕЯ).

Следует отметить, что выделение класса Семантической Эквивалентности (СЭ) высказывания является *важнейшей составляющей* любой задачи компьютерного анализа его смысла. В первую очередь, это обусловлено наличием синонимии как неотъемлемого свойства ЕЯ и, как следствие, возможностью выражения одного и того же смысла более чем одним способом. В общих чертах установить факт СЭ означает доказать идентичность ролей, в которых фигурируют идентичные понятия относительно сходных ситуаций, описываемых сравниваемыми текстами.

Поставим задачу установления СЭ ЕЯ-текстов следующим образом (*Плакат 2*). Пусть дано множество текстов G . Элементами G могут быть, к примеру, развернутые ответы обучаемых на вопрос тестирующей системы при применении заданий открытой формы. Требуется: по результатам синтаксического разбора исходных текстов выявить для каждого текста T_i :

- множество $V(T_i)$ ситуаций, описываемых текстом T_i ;
- множество $M(T_i)$ объектов (понятий), значимых в выявленных ситуациях;
- тернарное отношение I , которое ставит в соответствие каждому объекту ситуацию, в которой он фигурирует относительно заданного текста.

Далее на основе выявленного отношения необходимо выделить группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях. Иными словами, *имеем задачу* семантической кластеризации исходного множества текстов.

Идея предлагаемого решения основана на зависимости лексической сочетаемости слова от его Семантического Класа (СК) в заданном ЕЯ. С СК отождествляется обозначаемое словом понятие (сущность, предмет, явление) реального мира. Поэтому справедливо предположение о возможности выявления СК слова анализом его сочетаний с другими словами в ЕЯ-текстах по тематике заданной Предметной Области.

Следует отметить, что для извлечения СК слова из набора текстов заданной тематики первостепенную роль играет (*Плакат 3*) контекст целевого слова. Наибольшую точность, как показывает практика, дают модели контекста на основе синтаксических связей в предложении. В частности, для предикатных слов контекст определяется, в первую очередь, синтаксическими связями между предикатом и его семантическими актантами. Для формализации понятий, обозначающих участников тех или иных ситуаций, необходимо анализировать сочетаемость соответствующих существительных со словами, являющимися синтаксически главными по отношению к ним. Причем наряду с сочетаниями "актант–предикат" требуется учитывать произвольные сочетания существительных в тексте между собой (в том числе посредством предлогов).

Каждое выявляемое из текста понятие идентифицируется (в первую очередь) относительно заданного множества ситуаций. Сами ситуации обозначаются предикатными словами — глаголами (либо их производными). Поэтому наиболее приемлемым вариантом контекста для существительного, обозначающего некоторое понятие относительно анализируемой ситуации, будет представленная на *Плакате 3* последовательность из предиката и соподчиненных друг другу существительных. При этом значимым является тип отношения синтаксического подчинения между словами указанной последовательности, в первую очередь — для первого слова, обозначающего анализируемую ситуацию. В настоящей работе мы вводим в рассмотрение отношение R_q между произвольным словом указанной последовательности и ее последним словом. Реальные тексты Естественных Языков, в частности, русского, обладают тем свойством, что при наличии отношения R_q между первым и вторым словами в последовательности (1) на *Плакате 3* возможно установление данного отношения между первым и любым последующим словом вне зависимости от уже существующих отношений подчинения между словами этой последовательности. Данное свойство следует из соотношения смыслов соподчиненных слов. В настоящей работе мы будем использовать представленную на *Плакате 3* модель ситуационного контекста существительного как основу выявления понятий и ситуаций, в которых данные понятия фигурируют. С учетом указанного выше свойства отношения R_q рассматриваемые нами множества понятий и ситуаций расширяются в соответствии с *Алгоритмом 1 (Плакат 4)*. При этом роль, в которой объект выступает относительно некоторой ситуации, определяется типом q отношения R_q между словом, обозначающим ситуацию, и словом справа от него в последовательности соподчиненных слов. q характеризуется падежом зависимого слова и предлогом для связи главного и зависимого слова.

В качестве инструмента концептуальной кластеризации текстов в настоящей работе используются методы теории Анализа Формальных Понятий (АФП) — расширения теории решеток. При этом (*Плакат 5*) отношению I ставится в соответствие формальный контекст K и строится решетка Формальных Понятий \mathfrak{R} (*Плакат 5*) для исходного множества текстов G , а задача анализа смысловой близости текстов из множества G сводится к исследованию качественных характеристик решетки \mathfrak{R} .

Основные этапы построения решетки \mathfrak{R} представлены на *Плакате 6*. Данная решетка дает классификацию текстов относительно описываемых в них ситуаций и фигурирующих в этих ситуациях объектов. Визуализация решетки диаграммой линий позволяет графически отображать группировку текстов множества G по признакам вида "объект–ситуация–роль".

Тем не менее, при формировании множеств объектов и ситуаций на основе синтаксического анализа исходного множества текстов *актуальна проблема* наличия Расщепленных Значений (РЗ) в составе последовательностей соподчиненных слов. В настоящей работе за основу механизма выявления РЗ взяты правила синонимических преобразований ЕЯ-высказываний типа замещения с расщеплением в рамках стандартных Лексических Функций. Из известных видов РЗ наибольший практический интерес представляют Расщепленные Предикатные Значения (РПЗ). Формальное определение двух

основных известных в лексической семантике случаев РПЗ приведено на *Плакате 7*. При этом с учетом возможного наличия конверсивов применительно к обоим представленным на *Плакате 7* случаям РПЗ предполагается, что соответствующая замена уже выполнена, а совокупности S_1 и S_2 последовательностей соподчиненных слов, на основе которых задается РПЗ, описывают одно и то же множество объектов относительно одной и той же ситуации, обозначаемой словом v_{21} , то есть без мены ролей.

Определяемые *Утверждением 2* РПЗ включают в себя расщепления с глаголом-связкой, а также расщепления с глаголами—синтаксическими оформителями ситуаций, обозначаемых именами существительными, и представляющими собой языковое обозначение ролей участников ситуаций. *Актуальной* здесь является автоматическая лингвистически интерпретируемая классификация выявляемых РПЗ.

На *Плакате 8* представлено решение указанной задачи на основе методов АФП. Здесь вводится в рассмотрение формальный контекст K^{SV} для выявляемых РПЗ. При этом множество слов, являющихся синтаксически зависимыми в тех или иных РПЗ, рассматривается как множество G^{SV} формальных объектов, множество M^{SV} синтаксически главных слов — как множество формальных признаков для объектов из G^{SV} . Отношение I^{SV} ставит в соответствие каждому зависимому слову некоторого РПЗ множество слов, синтаксически главных по отношению к нему в тех или иных РПЗ.

Построение решетки Формальных Понятий для совокупности РПЗ представлено на *Плакатах 9–14*. Вначале на основе групп подряд идущих последовательностей соподчиненных слов на выходе синтаксического анализа *Алгоритмом 3 (Плакат 9)* выявляются пары соподчиненных слов, задающих РПЗ. Этим же алгоритмом производится замена найденных РПЗ на их однословные выражения во всех исходных последовательностях соподчиненных слов. Для выявленных РПЗ формируется описание в виде множества объектов с наборами признаков. Таким образом представляются кандидаты на включение в состав отношения I^{SV} .

Качественным показателем иерархичности формируемого ресурса является представленный на *Плакате 10* критерий полезности (4) для создаваемой решетки. С учетом требований данного критерия формирование решетки ведется по областям, то есть наборами ФП, связанных отношением порядка с одним Наибольшим Общим Подпонятием и/или одним Наименьшим Общим Суперпонятием. При этом в процессе генерации формального контекста пары (v_{12}, v_{11}) выбираются таким образом, чтобы всякое ФП, включаемое в решетку, входило в цепочку максимальной длины при максимизации его объема.

Формирование отдельной цепочки $P_{Ch(j)}^C$ на основе множества P^C объектов с заданными наборами признаков ведется согласно *Алгоритму 4 (Плакат 11)*. С целью минимизации числа спорных ФП, принадлежащих более чем к одной области с Наименьшими Общими Суперпонятиями, между которыми не существует отношение порядка, каждое следующее ФП в цепочке выбирается по принципу постепенного уменьшения содержания и максимизации количества общих признаков с потенциальным подпонятием при минимуме общих признаков с любым ФП, не входящим в цепочку.

Алгоритмом 5 (Плакат 12) строятся цепочки для Формальных Понятий, соседних по отношению к тем, между которыми устанавливается отношение порядка при формировании цепочки $P_{Ch(j)}^C$ Алгоритмом 4.

Максимум полезности (4) решетки достигается удалением наименее информативных признаков $v_{11} \in M^{SV}$ с наибольшими значениями частоты $Cnt(v_{11})$ встречаемости с различными $v_{12} \in G^{SV}$ из первоначально выявленных для $\{v_{12}\}$. Данная частота подсчитывается в соответствии с Алгоритмом 6 (Плакат 13) как число соответствующих употреблений $v_{12} \in M^{SV}$ в тексте. Максимизация полезности решетки удалением наименее информативных признаков из содержания всех Формальных Понятий во всех цепочках на выходе Алгоритма 5 осуществляется представленным на Плакате 14 Алгоритмом 7.

Для апробации предложенных алгоритмов был разработан программный комплекс, схема обмена данными между модулями которого представлена на Плакате 15. Синтаксический анализ осуществляется программой "Cognitive Dwarf" (ООО "Когнитивные технологии"). При тестировании данная программа показала самые точные результаты разбора.

Извлечение РПЗ из синтаксического дерева выделением последовательностей соподчиненных слов выполняет модуль *SV_revealing*. За основу при его реализации взята программа "Dwarfprint" в составе "Cognitive Dwarf". Генерацию контекста K^{SV} в соответствии с Алгоритмом 6 осуществляет разработанная авторами программа *XML_making*, которая представляет контекст K^{SV} на выходе Алгоритма 7 в виде XML-файла. С этой целью в программе *XML_making* реализована процедура индексирования признаков из M^{SV} . Визуализацию решетки диаграммой линий выполняет ПО *Concept Explorer*, реализующее методы АФП.

В качестве экспериментального текстового материала были взяты варианты ответов на тестовые задания открытой формы по материалам научных статей по тематике заданной Предметной Области. Используемое множество статей представляет собой тематическое подмножество того корпуса текстов, который по жанровому разнообразию представленного в нем рода словесности относится к научной прозе. Представленная на Плакате 16 решетка Формальных Понятий для выявленных РПЗ относится к тесту по материалам статьи К. В. Воронцова (ВЦ РАН), опубликованная в журнале "Таврический вестник информатики и математики", №1, 2004 г.

Основные результаты (Плакат 17) настоящей работы — алгоритм выявления и метод систематизации РПЗ в ЕЯ-текстах по тематике заданной Предметной Области. Следует отметить, что в настоящей работе рассмотрение ведется относительно последовательностей, которые состоят из глаголов (включая их особые формы — причастия и деепричастия) и существительных. Важнейшим направлением дальнейших исследований здесь является включение в состав указанных последовательностей наречий как характеристик действий, обозначаемых глаголами, и прилагательных как дополнительных характеристик объектов, обозначаемых существительными. Это позволит, в частности, учитывать расщепления с оценочными адъюнктами и расщепления на основе синтаксической деривации.