Формирование и кластеризация понятий на основе множества ситуационных контекстов.

Д. В. Михайлов, Г. М. Емельянов

Новгородский Государственный Университет имени Ярослава Мудрого

Настоящая работа посвящается *(Плакат 1)* решению *проблемы* адекватности формируемых знаний в процессе машинного обучения распознаванию Семантической Эквивалентности (СЭ) высказываний Естественного Языка (ЕЯ).

Выделение класса 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. Наиболее известная на сегодняшний день система классов СЭ в ЕЯ определяется множеством правил синонимических преобразований ЕЯ-высказываний в рамках стандартных Лексических Функций. В общем случае указанная система знаний строится на основе независимых текстовых описаний ситуаций (явлений) действительности выделением различающихся описаний одной и той же ситуации. Далее *ставится задача* установления степени близости между синонимичными описаниями, но уже различных ситуаций и формирования знаний о синонимии в виде прецедентов СЭ. Данная задача есть классическая задача *Распознавания Образов*.

При независимом ЕЯ-описании одной и той же ситуации возникает *проблема оценки адекватности* как формируемых *знаний* о синонимии, так и самих текстовых описаний ситуаций. Данная *проблема* актуальна для *частичных СЭ*. Наиболее *естественный путь* ее *решения* — выделение и систематизация понятий, значимых в описываемых ситуациях, непосредственно на основе СЭ-текстов заданной тематики с последующей качественной оценкой формируемой понятийной структуры тезауруса. Разработка математической модели указанного процесса является *целью* настоящей работы *(Плакат 1)*.

Идея предлагаемого решения основана на зависимости лексической сочетаемости слова от его Семантического Класса (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 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. Поскольку сами ситуации обозначаются предикатными словами — глаголами (либо их производными), наиболее преемлемым вариантом контекста для существительного, обозначающего некоторое понятие относительно анализируемой ситуации, будет представленная на *Плакате 2* последовательность слов, состоящая из предикатного слова и соподчиненных друг другу существительных. При этом значимым является тип отношения R_q 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, обозначающим выявляемое понятие. При наличии такого рода сочетаний в анализируемом тексте мы можем рассуждать как о правильности ЕЯ-описания рассматриваемых контекстов носителем предметных знаний — человеком, так и о частичных СЭ, задаваемых относительно последнего слова в (1) последовательностями вида (2) *(Плакат 2)*. Сказанное подтверждается тем фактом, что реальные тексты Естественных Языков, в частности, русского, обладают тем свойством, что при наличии отношения R_q между первым и вторым словами в последовательности (1) возможно установление данного отношения между первым и любым последующим словом в последовательности (1) вне зависимости от уже существующих отношений подчинения между словами этой последовательности. Данное свойство следует из соотношения смыслов соподчиненных слов. Ярким примером могут послужить Генитивные Конструкции русского языка, которые являются частным случаем Семантической Эквивалентности, определяемой *Утверждением 1* на *Плакате 2*. Сравнение (*Плакат 2*) : *"сложность подсемейства модели" — "сложность модели"*, *"Характеристика сложности семейства алгоритмов" — "характеристика алгоритмов"*.

Сказанное позволяет ассоциировать с последовательностями вида (2) более абстрактные ситуации, чем с (1), а саму последовательность (1) рассматривать в качестве ситуационного контекста для существительного, обозначающего выявляемое понятие. В настоящей работе в качестве базовой структуры для выявления и кластеризации понятий предлагается использовать ситуационные контексты (1), которые участвуют в описании частичных СЭ в соответствии с *Утверждением 1*. Ставится *задача* : по результатам синтаксического разбора выявить указанные контексты в анализируемом тексте и на их основе выполнить концептуальную кластеризацию.

Результатом синтаксического анализа является набор деревьев разбора предложений. Далее с каждого дерева последовательно считываются пары *"синтаксически главное слово — зависимое слово"*. Дальнейшая обработка считанных пар направлена на выявление последовательностей вида (1) и (2), удовлетворяющих условию *Утверждения 1*. Как результат формируется множество ситуационных контекстов — последовательностей вида (1).

В качестве инструмента концептуальной кластеризации ситуационных контекстов в настоящей работе используются методы теории Анализа Формальных Понятий (АФП) — расширения теории решеток. При извлечении из текста множество существительных, обозначающих выделяемые понятия, рас-

сматривается в качестве множества объектов G *(Плакат 3)*. В множество признаков M включаются существительные и глаголы, которые входят в состав последовательностей вида (1), участвующих в описании частичных СЭ относительно слов из G в соответствии с *Утверждением 1*. Отношение I ставит в соответствие каждому существительному из множества G те соподчиненные слова некоторой последовательности вида (1), для которых в анализируемом тексте присутствуют последовательности вида (2) относительно заданного $g_k \in G$. В множестве M слова представлены вместе с обязательными предлогами, посредством которых синтаксически главное слово $m \in M$ связывается с зависимым словом из G. Тройка $K = (G, M, I)$ называется формальным контекстом, а множество всех Формальных Понятий (ФП) контекста вместе с заданным на нем отношением порядка — решеткой ФП.

Построение модели тезауруса в виде решетки ФП представлено на *Плакатах 4 — 9*. Вначале с учетом свойства отношения R_q между первым и вторым словами последовательности (1) *Алгоритмом 1 (Плакат 4)* формируются пары-кандидаты на включение в состав отношения I.

Качественным показателем иерархичности формируемого тезауруса является представленный на *Плакате 5* 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 (g_k, m) выбираются таким образом, чтобы всякое ФП, включаемое в решетку, входило в цепочку максимальной длины при максимизации его объема.

Формирование отдельной цепочки $P^C_{Ch(j)}$ на основе множества P^C объектов с заданными наборами признаков ведется согласно *Алгоритму 2 (Плакат 6)*. 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 минимальном количестве общих признаков с любым ФП, не входящим в цепочку.

Алгоритмом 4 (Плакат 7) строятся цепочки для Формальных Понятий, соседних по отношению к тем, между которыми устанавливается отношение порядка при формировании цепочки $P^C_{Ch(j)}$ *Алгоритмом 2*.

Максимум полезности (4) решетки достигается удалением наименее информативных признаков $m \in M$ с наибольшими значениями частоты $Cnt(m)$ встречаемости с различными $g_k \in G$ из первоначально выявленных для $\{g_k\}$. Данная частота подсчитывается в соответствии с *Алгоритмом 5 (Плакат 8)* как число соответствующих употреблений $m \in M$ в тексте. Максимизация полезности решетки удалением наименее информативных признаков из содержания всех Формальных Понятий во всех цепочках на выходе *Алгоритма 4* осуществляется представленным на *Плакате 9 Алгоритмом 6*.

Для апробации предложенных алгоритмов был разработан программный комплекс, схема обмена данными между модулями которого представлена на *Плакате 10*. Синтаксический анализ осуществляется программой *"Cognitive*

Dwarf“ (ООО ”Когнитивные технологии“). При тестировании данная программа показала самые точные результаты разбора.

Извлечение потенциальных пар (g_k, m) из синтаксического дерева выполняет модуль *Pairs*. За основу при его реализации была взята программа *”Dwarfprint“* в составе *”Cognitive Dwarf“*. Генерацию контекста $K = (G, M, I)$ в соответствии с *Алгоритмом 6* осуществляет разработанная авторами программа *XML_ making*, которая представляет контекст K на выходе *Алгоритма 6* в виде XML-файла. С этой целью в программе *XML_ making* реализована процедура индексирования признаков из M. Визуализацию решетки диаграммой линий выполняет ПО *Concept Explorer*, реализующее методы АФП.

Исходными данными для формирования решетки понятий являются рефераты 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. Основным требованием к используемому множеству текстов является репрезентативность. Под репрезентативностью *(Плакат 11)* в настоящей работе понимается способность множества текстов отображать все свойства Предметной Области, релевантные проводимому исследованию. За оценку репрезентативности в работе принято отношение суммарной частоты встречаемости последовательностей вида (1) к количеству выявленных в них типов синтаксических отношений.

В качестве экспериментального текстового материала была обзорная статья К. В. Воронцова (ВЦ РАН), опубликованная в журнале ”Таврический вестник информатики и математики“, №1, 2004 г. Указанная публикация является хорошим примером репрезентативного текста в соответствии с представленным на *Плакате 11* критерием с характерной минимизацией числа типов отношений для последовательностей соподчиненных слов при максимизации количества таких последовательностей. Полученная для статьи К. В. Воронцова решетка Формальных Понятий представлена на *Плакате 12*.

Основной *результат (Плакат 13)* настоящей работы — *алгоритм формирования понятийной структуры тезауруса Предметной Области* на основе описывающих ее ЕЯ-текстов. Предложенная в работе *модель тезауруса в виде решетки ФП* позволяет оперировать данными на семантическом уровне без потери или недопустимого упрощения объектов и их признаков.

Сфера применения полученных результатов — автоматизация пополнения лингвистических информационных ресурсов. Пример — тезаурус по анализу изображений, разрабатываемый ВЦ РАН.

Наибольшая эффективность *предложенного метода формирования и кластеризации понятий* достигается при совместном его использовании с анализом сочетаемости предикатных слов в рамках семантико-синтаксических валентностей. Здесь *перспективным направлением дальнейших исследований* следует указать *(Плакат 14)* развитие предложенного метода применительно к Расщепленным Значениям в составе соподчиненных слов.

С учетом результатов машинного эксперимента *отдельного рассмотрения заслуживает* предварительная обработка текстового материала с максимизацией его репрезентативности. *Перспективным* здесь *является* выделение и замена анафор, в первую очередь — анафорических личных местоимений.