Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности.

Д. В. Михайлов, Г. М. Емельянов

Новгородский Государственный Университет имени Ярослава Мудрого

Настоящая работа посвящается *(плакат 1)* решению проблемы машинного обучения распознаванию ситуаций 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 и адекватная с лингвистической точки зрения система классов СЭ в ЕЯ определяется множеством правил синонимических преобразований ЕЯ-высказываний в рамках стандартных Лексических Функций (ЛФ). Указанные правила описывают Ситуации СЭ (ССЭ) на уровне варьирования универсальной (абстрактной) лексикой, что особенно актуально для реальных текстов. В большинстве случаев синонимия обусловлена именно варьированием абстрактными словами и их сочетаниями. Предметная лексика, как правило, остается без изменений.

Как известно, значительную трудность при практической реализации указанных преобразований представляет формализация особого компонента правила, именуемого условием его применимости. Это условие является совокупностью требований к синтаксическим и семантическим свойствам заменяемых лексических единиц и выполняет функцию фильтра. Если конечный продукт синтеза дает нарушение Лексического Значения (ЛЗ), сочетаемости или стилистических норм, фильтр отбраковывает синтез определенной фразы из множества семантически эквивалентных. Многие фильтры были сформулированы в работах И. А. Мельчука, И. А. Жолковского. Однако, как отметил академик Ю. Д. Апресян, проблема нуждается в дальнейшей разработке. Тем более, что по оценке И. А. Мельчука, специальных исследований по данному вопросу не проводилось, а сами правила описаны в первом приближении.

В содержательном плане условие применимости правила синонимического преобразования составляет основу прецедента класса СЭ высказываний *(плакат 2)*. Исходными данными при формировании прецедента являются признаки слов в парах сравниваемых по смыслу высказываний. Причем помимо уже известных классов СЭ, описанных в работах Московской лингвистической школы, в ходе анализа могут быть выделены и новые классы СЭ, выходящие за рамки лексической семантики. Данная задача есть классическая задача Распознавания образов.

Исходя из вышеизложенного, *цель* настоящей работы сформулирована как *(плакат 1) разработка и исследование методик формирования прецедентов для классов Семантической Эквивалентности в Естественном Языке.* При этом основной акцент внимания в работе делается классам СЭ на основе стандартных Лексических Функций.

Наибольший интерес для описания прецедента класса СЭ в нашей постановке задачи представляют ситуации с ЛФ-параметрами *(плакат 3)*. Посредством этих ЛФ описывается Расщепленное Значение (РЗ). При этом требования к заменяемым лексическим единицам исходного ЕЯ-высказывания в составе прецедента класса СЭ определяются Смысловыми Отношениями (СО) между некоторым ЕЯ-словом, относительно которого задается ССЭ (ключевым словом ЛФ-синонимической замены), и его лексико-семантическими производными (лексическими коррелятами), которые входят в заменяемый комплекс лексических единиц. В лексической семантике такие СО описываются стандартными ЛФ. Фактически СО непосредственно задается Расщепленным Значением. Это позволяет поставить задачу его выявления и обобщения по аналогии с описанием семантики Именных Групп на основе формализованного представления толкований Лексических Значений слов в виде теорий *(плакаты 2,4)*. Сказанное подтвержается наработками по Русскому общесемантическому словарю (РОСС): ЛФ используются в качестве Семантических Характеристик отдельных слов в РОСС. Следовательно, такие слова могут быть и названиями отношений в утверждениях теорий других слов. Пример — значение Лексической Функции $Oper_1$ для ЛЗ «эксперимент» (т. е. «осуществлять», *(плакаты 3,12)*). Как видно из *рис.5* на *плакате 12*, оно присутствует в одном из утверждений теории ЛЗ «экспериментировать». Данное ЛЗ эквивалентно РЗ «осуществлять эксперимент», где значением ЛФ $Oper_1$ задается СО типа «операция с. . . » между 1-м участником ССЭ (кто осуществляет эксперимент) и ее названием.

Согласно определению, теория Лексического Значения слова представляет собой набор утверждений (постулатов значения, *плакат 4*), посредством которых данное слово связывается с другими словами и понятиями. При этом смысл слова определяется набором Характеристических Функций (ХФ, *Определение 2, плакат 4*), которые задаются утверждениями теории ЛЗ слова и составляют для него набор признаков. Пример такого набора для ЛЗ «агрессор» приведен на *плакате 5*. При независимом построении теории одного слова разными исследователями возникает задача обобщения получаемых знаний. Сказанное актуально при построении теорий на основе ЕЯ-толкований слов с применением стандартных концептуальных языков.

Для решения показанной задачи *(задача 2, плакат 1)* в настоящей работе путем применения математических методов Анализа Формальных Понятий (АФП) и реализующего эти методы специализированного ПО ToscanaJ (http://toscanaj.sourceforge.net) строится представленная на *плакатах 6,7 и 8* модель системы элементов толкования для независимых теорий Лексического Значения заданного слова. Действительно, Лексическое Значение слова, описываемое посредством формализованной теории *(плакат 4)*, есть денотация. В логике ей ставится в соответствие экстенсионал как класс сущностей, которые определяются посредством теории. При этом внешне различные описания теорий одного и того же Лексического Значения определяют единое множество Характеристических Функций, задаваемых в соответствии с *Определением 2* на *Плакате 4*. Характеристические Функции (в том числе определяемые рекурсивно для списочных аргументов отношений произвольной арности, *Плакат 6*) задают набор формальных признаков для элемен-

тов толкования Лексического Значения. В конечном итоге они определяют интенсионал обобщенной теории заданного Лексического Значения. Таким образом, исходя из определения интенсионала как функции от возможных миров к экстенсионалам, а также рекурсивной природы постулатов значения, имеем задачу построения обобщенной теории Лексического Значения как восстановление синтаксического представления экстенсионала на основе известного синтаксиса λ-выражений для Характеристических Функций, составляющих интенсионал *(плакаты 9,10)*. На основе решетки Формальных Понятий для Лексического Значения *(Плакат 7)* ключевое правило обобщения утверждений независимых теорий Лексического Значения определяется введением в рассмотрение области, которую образуют элементы толкования заданного Лексического Значения в решетке *(плакат 9)*. Это позволяет различать случаи :

- использования разных Характеристических Функций с одним и тем же значением в независимых альтернативных вариантах теории Лексического Значения (обобщение посредством отношения «или»). В представленном на *Плакате 8* формальном контексте примерами могут послужить пара Формальных Понятий : *(«Толкование2_ агрессор», «Толкование3_ агрессор»)* и пара, образованная *«Толкованием1_ агрессор»* и НОСП для пары *(«Толкование2_ агрессор», «Толкование3_ агрессор»)*;
- описания одного и того же элемента толкования ЛЗ, но посредством разных Характеристических Функций (обобщение посредством отношения «и»). В представленном на *Плакате 8* формальном контексте примером может послужить содержание (совокупность формальных признаков) ФП *«Толкование1_ агрессор»*, а также содержанию НОСП для пары *(«Толкование2_ агрессор», «Толкование3_ агрессор»)*.

При этом вычислительная сложность процесса обобщения теорий заданного ЛЗ зависит исключительно от мощности множества Характеристических Функций. Согласно определению смысла как интенсионала Лексического Значения, последняя не зависит от числа обобщаемых теорий. В перспективе для утверждений, объединяемых посредством утверждения «или», здесь появляется возможность задействования статистических методов для выявления наиболее значимых признаков.

Для решения задачи выявления и обобщения Смыслового Отношения в рамках РЗ в настоящей работе проводится аналогия между описанием ЛЗ слова в виде теории и описанием Семантического Класса слова посредством дескрипторов таксономических категорий (КАТ) и Семантических Характеристик (СХ) в РОСС. Применение Лексических Функций в качестве СХ отдельных слов в данном словаре позволяет сделать вывод о возможности выявления смысловых зависимостей, определяемых Лексическими Функциями, путем сравнительного анализа множеств аксиом теорий ЛЗ слов в Расщепленном Значении *(плакат 11)*. Сравнение производится на предмет наличия зависимости, определяемой Семантическим Отношением в некотором утверждении вида (2) или (3) *(плакат 4)* одной из сопоставляемых теорий. При этом *(плакат 11)* подмножество аксиом теории ЛЗ другого слова либо является одним из аргументов этого отношения, либо непосредственно задается одним из сравниваемых слов. Примером могут послужить теории ЛЗ «экс-

перимент» и «экспериментировать» *(плакат 12)*.

Лексическими Функциями описывается в первую очередь лексическая сочетаемость, которая определяется Лексическим Значением ключевого слова ЛФ-синонимической замены. Следовательно, ЛЗ более узкого по смыслу слова (в терминологии АФП — гипонима) включает Лексические Значения более широких по смыслу слов (гиперонимов), которые упоминаются в толковании ЛЗ рассматриваемого слова, а, следовательно, и в его теории. Таким образом, слово-гипоним в большинстве случаев будет иметь в качестве значений ЛФ-параметра значения этой же ЛФ для тех слов-гиперонимов, которые упоминаются в его толковании (теории).

Сказанное позволяет для заданной ЛФ построить модель системы словее аргументов, представленную на *плакатах 13* и *14*. При этом *(плакат 13)* ключевые слова-аргументы заданной Лексической Функции выступают в качестве объектов, а слова-значения этой Лексической Функции — в качестве формальных признаков.

Само отношение гипонимии («подпонятие-суперпонятие») на множестве слов-аргументов заданной ЛФ в простейшем случае будет иметь место при наличии показанного на *плакате 15* взаимно-однозначного соответствия между множествами семантических классов актантов с идентичными ролевыми ориентациями у гипонима и гиперонима. Но с учетом возможности использования Лексических Функций-параметров в составе цепочек СХ отдельных слов в РОСС, следует расширить определяемое нами отношение гипонимии на множестве Лексических Значений предикатных слов введением в рассмотрение зависимостей между СХ актантов предикатных слов. При этом взаимно-однозначное соответствие между Семантическими Классами актанта гипонима и гиперонима устанавливается путем поиска общих подсписков Семантических Характеристик в совокупности с вхождением Семантических Характеристик одного актанта в утверждения теорий Семантических Характеристик другого актанта так, как показано на *плакате 16*.

Примером указанного соответствия может послужить аспектная валентность у ЛЗ «испытания» и валентность содержания у ЛЗ «тест» *(плакат 17)* для представленных на *плакате 14* слов из верхней окрестности ЛЗ «эксперимент». Действительно, согласно указанному в *Утверждении 4* на *плакате 15* условию существования отношения гипонимии между Лексическими Значениями, ЛЗ «тест» не может выступать в качестве суперпонятия для ЛЗ «испытание». Основание — отсутствие задаваемого *Утверждением 4* соответствия для валентности аспекта у ЛЗ «испытание» и валентности содержания у ЛЗ «тест». Тем не менее, в словарной Базе Данных АРМ лингвиста на (http://www.aot.ru), для Семантического Класса слова, реализующего аспектную валентность у ЛЗ «испытание» и для Семантического Класса слова, реализующего валентность содержания у ЛЗ «тест», имеются представленные на *плакате 17* описания совокупностями вышеупомянутых дескрипторов. Кроме того, имеем также теорию сорта, отождествляемого с СХ «SITUAT» *(рис.7, плакат 17)*. Как видно из приведенного на *плакате 17* древовидного описания, теория сорта «SITUAT», упоминаемого в списке СХ для ЛЗ «ситуация», «ссылается» на Семантические Характеристики «ATTR» и «PARAM», из которых «ATTR» присутствует в списке СХ для ЛЗ «свой-

ство». Таким образом, относительно ЛЗ «испытание», ЛЗ «тест» удовлетворяет сформулированным нами требованиям к суперпонятию Лексического Значения.

Путем визуализации *(плакат 18)* средствами Visual Prolog'а 5. 2 расширенного нами отношения гипонимии для множества Семантических Классов слов-аргументов заданной ЛФ мы можем оценить как адекватность и полноту описания слова по заданной ЛФ, так и корректность его лексикографического толкования как основы для построения Модели Управления этого слова *(плакат-приложение)*. При этом *(плакат 19)* справедливым будет утверждать, что теории пары ЛЗ адекватно описывают условие применимости правила ЛФ-синонимического преобразования ЕЯ-высказывания с расщепленным значением как основы формирования прецедента соответствующего класса СЭ при выполнении слудующего условия. Название выявляемого Смыслового Отношения в рамках Расщепленного Значения должно принадлежать множеству формальных признаков того ЛЗ, которое есть НОСП для множества слов верхней окрестности слова, выполняющего в Расщепленном Значении функцию аргумента выявляемого Смыслового Отношения. При этом НОСП определяется применительно к отношению гипонимии на множестве Семантических Классов слов-аргументов заданной ЛФ *(плакат 15 и 16)*.

Разработанная методика формирования прецедентов для классов СЭ, определяемых на основе Расщепленных Значений с Лексическими Функциями-параметрами, была апробирована на материале лексикографических толкований слов окрестности ЛЗ «эксперимент», а также независимых толкований ЛЗ «агрессор». Варианты толкований взяты из Большой Советской Энциклопедии, тематического словаря «Война и мир» и словаря Брокгауза и Ефрона на http://slovari.yandex.ru а также из Толково-комбинаторного словаря современного русского языка И. А. Мельчука и А. К. Жолковского.

Тем не менее, следует отметить, что используемое в настоящей работе описание смысла слова набором Характеристических Функций производится в шкале наименований. При обобщении утверждений независимых теорий одного и того же Лексического Значения посредством отношения «или» не учитывается статистическая значимость каждого признака. Значения Характеристических Функций, которые задаются объединяемыми утверждениями, полагаются равновероятными. В связи с этим в качестве перспективного направления дальнейших исследований следует выделить введение в рассмотрение распределений возможных значений Характеристических Функций как формальных признаков Лексического Значения. Это позволит вычислять меру близости между предикатами, описывающими теории ЛЗ слов в рамках Расщепленного Значения и тем самым сократить объем обучающей выборки при формировании прецедента класса СЭ.

Задействование Характеристических Функций при описании смысла слова и их выводимость из теории его Лексического Значения позволяет в перспективе ввести в рассмотрение родовидовые зависимости между теориями при описании прецедентов для ситуаций СЭ на основе неточных синонимов, конверсивов и дериватов. Это позволит фиксировать различия в актантной структуре этих слов.