Применение сортовой системы Естественного Языка в задаче установления смысловой эквивалентности текстов.

Д.В.Михайлов, Г.М.Емельянов
Новгородский государственный университет имени Ярослава Мудрого

Настоящая работа посвящена (*плакат 1*) проблеме *учета контекста синонимического преобразования Глубинной Синтаксической Структуры (ГСС) при решении задачи установления смысловой (семантической) эквивалентности высказываний Естественного Языка (ЕЯ) в рамках теоретического подхода "Смысл⇔Текст"*.

К числу наиболее актуальных практических задач, требующих установления смысловой эквивалентности ЕЯ-текстов, относится интерпретация результатов открытых тестов в системах компьютерного дистанционного обучения и контроля знаний (*плакат 2*). Эта задача нами уже обсуждалась на предыдущих конференциях, в частности, *5-й Международной научной конференции "Интеллектуализация обработки информации" (ИОИ-2004), Алушта, 2004 и 12-й Всероссийской конференции Математические методы распознавания образов (ММРО-12), Москва, 2005*.

Применение языка глубинного синтаксиса в качестве языка смыслов в рамках теоретического подхода к языку как преобразователю "Смысл⇔Текст" дает возможность использования для сравнения смыслов высказываний конечного числа корректно формализуемых преобразований помеченных деревьев (*плакат 3*). Основным достоинством такого подхода является независимость правил синонимических преобразований от предметной области высказываний. Указанные правила синонимических преобразований описывают ситуации лексико-синтаксических замен на уровне варьирования универсальной (абстрактной) лексикой в рамках аппарата стандартных Лексических Функций (ЛФ), что особенно актуально для реальных тестов : в большинстве случаев обучаемый употребляет синонимы именно на уровне абстрактных слов и их сочетаний, оставляя предметную лексику без изменений (*плакат 3*).

Тем не менее, *значительную трудность* при практической реализации указанных преобразований в программных системах анализа смысловой эквивалентности высказываний *представляет формализация* особого компонента *правила*, именуемого *условием* его *применимости*. В содержательном плане условие применимости лексического правила представляет собой совокупность требований к синтаксическим и семантическим свойствам лексических единиц исходной ГСС, входящих в заменяемое правилом поддерево. Подобные ограничения отражают особенности лексики конкретного ЕЯ и выполняют функции фильтров, задерживающих синтез определенной фразы из множества семантически эквивалентных, если конечный продукт синтеза дает нарушение Лексического Значения (ЛЗ), сочетаемости или стилистических норм. 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 синонимических преобразований ГСС с применением аппарата стандартных ЛФ описаны в первом приближении.

Следует отметить, что значительное число изложенных, в частности, Ю.Д. Апресяном требований к семантике и синтаксису лексических единиц, фигурирующих в правилах, относится к соответствующим способам реализации семантических валентностей либо ключевого слова синонимического преобразования, либо лексических коррелятов этого слова, входящих в заменяемый правилом комплекс лексических единиц. Подобного рода "отсеивающие" фильтры использовались и используются, в частности, для выбора правильного варианта разбора предложения в программах синтаксического анализа. Важнейшую роль при этом играет информация семантической компоненты Модели Управления (МУ) предикатного слова, ее суть была рассмотрена нами на предыдущей конференции *ММРО-12*.

Тем не менее, в ряде случаев простого указания Семантического Класса (СК), которым должен быть выражен тот или иной актант предикатного слова, для адекватного выбора варианта анализа (перифразирования) предложения оказывается недостаточным. Причина здесь заключается в наличии для одного предикатного слова нескольких Лексических Значений (ЛЗ), с каждым из которых связывается альтернативный вариант МУ и соответствующий синоним с более широким, чем у самого слова, значением. Примером может послужить глагол *сжечь (плакат 4)*, который может вне специального контекста в равной мере быть интерпретирован и как *уничтожить*, и как *израсходовать* (подобные случаи неполных смысловых эквивалентностей описаны также И.А. Мельчуком в *"Опыте теории лингвистических моделей "Смысл⇔Текст""*, смотри пример с *заплатил* на стр. 152). При использовании системы упомянутых выше правил синонимических правил синонимических преобразований ГСС названный фактор (теоретически) может привести к построению неадекватных перифраз *(плакат 4)*. Данная проблема в ряде случаев может быть решена, как показано в работах академика Ю.Д. Апресяна, наложением частных ограничений на применимость конкретных правил. Эти ограничения касаются грамматических и семантических зависимостей, отличных от актантных. Путем неприменимости правила, например, в случае наличия у исходного (ключевого) слова адъюнкта определенного СК *(плакат 5)* делается соответствующий вывод о конкретном ЛЗ из множества возможных для данного ключевого слова. Тем не менее, подобного рода фильтры работают не всегда : наличие адъюнкта у ключевого слова не является обязательным и оба ЛЗ становятся в равной мере возможными.

Наиболее естественный путь решения показанной проблемы заключается в привлечении информации словарных определений (дефиниций) обозначаемых актантами предикатного слова понятий. При этом введение одновременно в рассмотрение подобных определений для

семантики отношений, отличных от связей предиката с актантами по МУ и задаваемых входящими в анализируемое предложение Именными Группами (ИГ), позволяет более точно устанавливать соответствия требованиям семантической интерпретации глубинных синтаксических актантов предикатного слова. Данная точка зрения близка к сравнительно недавно возникшему направлению в вычислительной семантике (*computational semantics*), которое в публикациях отечественных авторов характеризуется как структурное изучение лексической и контекстной семантики. В рамках этого направления следует выделить работы по приложению предложенного американским логиком *Ричардом Монтегю* языка интенсиональной логики для формализации описания семантики *Именных Групп*. Как было показано в исследованиях Борщева В.Б. и Б.Х. Парти, используемый в интенсиональной логике оператор лямбда-абстракции может быть успешно использован для формального описания Значения слова (фактически – Лексического Значения слова) в виде набора утверждений, связывающих его с другими словами и понятиями. При этом употребляемое Б.Х. Парти и В.Б. Борщевым понятие сорта как элемента "наивной картины мира" и класса, к которому язык относит конкретную реалию, фактически соответствует тому, что в публикациях Московской лингвистической школы понимается под СК обозначающего эту реалию слова. Это же понимание СК используется нами применительно к описанию семантической интерпретации глубинного синтаксического актанта предикатного слова.

Исходя из вышеизложенного, *цель* настоящей работы сформулирована как (*плакат 1*) *исследование практических аспектов построения теории ЛЗ слова по его словарному определению для автоматизации выделения свойств обозначаемого словом объекта и формирования в полуавтоматическом режиме множества аксиом теории*.

Для достижения поставленной цели были сформулированы следующие задачи :

- Разработать реализуемое программно формальное описание теории ЛЗ слова с ориентацией на общие принципы использования семантической информации при построении дерева синтаксического подчинения и полученную ранее структуру языковой базы знаний для задачи установления семантической эквивалентности текстов ЕЯ;
- На основе введенного формального представления разработать методику систематизации и контроля корректности наборов утверждений-аксиом, используемых при построении теорий.

При решении *первой* из поставленных *задач* в качестве основного требования к формальному описанию теории ЛЗ в виде принятого в формальной семантике λ-выражения была выдвинута реализуемость динамической базы процедурных знаний средствами традиционных типизированных языков программирования. Были сформулированы представленные на *плакате 6* свойства подобного описания значения слова, актуальные для его программной реализации с учетом указанного требования и применения информации разработанной нами ранее языковой базы знаний

при построении аналогов λ-выражений самим пользователем “на лету” с последующим добавлением в базу знаний системы непосредственно во время работы. Среди показанных свойств в качестве наиболее существенного для решаемых в настоящей работе задач следует отметить *возможности типовых и сортовых сдвигов*, позволяющие ввести в рассмотрение теории для отношений, задаваемых опорными словами ИГ и указываемыми в качестве семантической ориентации актантов предикатных слов. С учетом синтаксиса λ-выражений, принятого в формальной семантике, а также представленных на *плакате 6* свойств рассматриваемого описания Лексических Значений слов, теория ЛЗ слова, а также теория задаваемого им отношения могут быть, в частности, описаны представленными на *плакатах 7 и 8* составными объектами языка Пролог.

При наличии формализованного описания теорий Лексических Значений (в виде утверждений (6, *плакат 9*) динамической базы фактов Visual Prolog’a) и задаваемых ими отношений (в виде утверждений вида (7, *плакат 9*)) определение принадлежности слова сорту (то есть СК) с заданной теорией, а также автоматическое выявление отношения, задаваемого ИГ (на основе данных о Семантических Классах входящих в нее слов), организуется как сопоставление (с предварительной конкретизацией переменных в составных объектах *(1*, *плакат 7)* и *(4*, *плакат 8)*) аксиом из списков *Rel_list* с занесенными в динамическую базу фактов утверждениями *relation2* в соответствии с приведенными на *плакате 9* Пролог-правилами.

Располагая информацией об известных видах реляционных ИГ на основе выявления морфологических характеристик и Семантических Классов слов, входящих в заданную ИГ, можно организовать распознавание интересующего нас отношения непосредственно в тексте по факту присутствия ИГ, которая это отношение задает. Решение указанной задачи может быть описано представленным на *плакате 10* Пролог-правилом.

Процесс подготовки для последующего занесения в базу знаний формализованных описаний теорий Лексических Значений и задаваемых ими отношений в простейшем случае можно организовать как выбор соответствующих значений (имен сортов, названий отношений, их аргументов, обозначений для переменных) из предлагаемых экранной формой списков. При этом составные объекты, которые используются для представления указанных теорий, могут быть легко преобразованы в структуры, принадлежащие определенному для работы с деревьями в Visual Prolog'e специальному домену tree. Это позволяет задействовать средства Visual Programming Interface (VPI) для контрольного вывода на экран древовидного описания теорий Лексического Значения *(плакат 11)* и задаваемого им отношения *(плакат 12)*.

Тем не менее, визуальный контроль наборов утверждений-аксиом с помощью представленных на *плакатах 11* и *12* экранных форм оказывается недостаточным для анализа корректности как используемых наборов, так и содержательной интерпретации аксиом для независимого описания теорий разных ЛЗ. Причина заключается в *возможности непустого пересечения*

множеств аксиом теорий у *описываемого* и *одного из известных* сортов. Примером может послужить ситуация, когда слово одновременно следует отнести к нескольким сортам : в этом случае его теория должна содержать аксиомы, специфичные для каждого из этих сортов плюс дополнительные аксиомы, отражающие специфику ЛЗ самого слова. Здесь же возникает техническая проблема избыточности информации Базы Данных, представляемой совокупностью утверждений вида *(6)* и *(7)*.

Для решения указанной проблемы в настоящей работе мы используем свойства онтологической классификации, лежащей в основе деления на сорта внутри понятийной системы.

Действительно, поскольку теория слова “ссылается” к теориям всех сортов, к которым слово принадлежит, то при использовании предложенного в настоящей работе подхода к формальному описанию теорий аксиомы той части теории слова, которая является "отсылкой" к теории соответствующего сорта, будут находиться во взаимно-однозначном соответствии с аксиомами теории самого сорта. При этом *внутри* каждой *теории* выделяются *сходные группы аксиом*, и *каждая из* этих групп задает свою *функцию* таким образом, что значения идентичных функций у теорий слов с более узким и более широким смыслом будут связаны отношением "*род-вид*", за счет чего принадлежащие заданному сорту слова как объекты образуют иерархию. Сказанное позволяет применить к материалу Базы Данных теорий Лексических Значений широко известные за рубежом математические методы Формального Концептуального Анализа (ФКА).

Для системы теорий слов имеем представленный на *плакате 13* частный случай многозначного контекста, в котором отношение *“субконцепт-суперконцепт”* задается путем вычисления значений функций, однозначно характеризующих ключевые свойства обозначаемых словами объектов реальности. Примеры реализующих такие функции *λ-выражений* для вычисления названий-этикеток характеристики, классифицируемой как *“объем”*, той части некоторого физического объекта, которую можно рассматривать как *“полость”*, приведены на *плакате 14*. Для построения и визуализации формального контекста в настоящей работе было использовано реализующее методы ФКА специализированное ПО ToscanaJ (http://toscanaj.sourceforge.net/). Визуальное представление *(плакат 15)* формального контекста для системы теорий слов-представителей заданного сорта позволяет оценить степень сходства и различия теорий отдельных слов внутри сорта, на основе чего можно сделать вывод об адекватности набора аксиом, задающих теорию самого сорта. *Перспективным* здесь *является* изучение возможностей построения формализованных теорий для контекстных отношений, не задаваемых напрямую ЛЗ опорного существительного ИГ.

В качестве *перспективного направления исследований* следует также отметить разработку модели самого правила синонимического преобразования ГСС с ориентацией на предложенную в настоящей работе методику представления и использования информации о ЛЗ слова.