Морфология и синтаксис в задаче семантической кластеризации.

Михайлов Д. В., Емельянов Г. М.

Новгородский Государственный Университет имени Ярослава Мудрого

Актуальная *глобальная задача*, которой посвящена настоящая работа — автоматизация накопления знаний о взаимодействии семантики, синтаксиса и морфологии при установлении Семантической Эквивалентности (СЭ) текстов Естественного Языка (ЕЯ).

Выделение класса СЭ высказывания является *важнейшей составляющей* любой задачи компьютерного анализа его смысла. В общих чертах установить факт СЭ означает доказать идентичность ролей сходных понятий относительно сходных ситуаций, описываемых сравниваемыми текстами. Наиболее близка данной идее обработка текстов на основе коммуникативной грамматики. Хорошим примером является поисковая система Exactus.

Тем не менее, существуют задачи сравнения смысла, отличные от традиционного для поисковых систем взаимодействия «запрос–ответ». Примером является тестовое задание открытой формы в системе контроля знаний. Необходимо не столько отобразить ответ на предметную область, сколько оценить близость ответу, «правильному» с точки зрения разработчика теста. Анализ взаимной близости ответов здесь требует учета лексико-функциональной синонимии, в частности — расщепленных значений и конверсивов. По оценке Г. С. Осипова, необходимо более объективное исследование свойств семантических связей и в самой коммуникативной грамматике. Актуальным является задействование методов машинного обучения для автоматизации накопления знаний о синонимии в ЕЯ.

Наиболее естественная постановка задачи СЭ с учетом указанных требований к процессу сравнения текстов состоит в следующем *(Плакат 2)*. Пусть дано множество текстов. Элементами этого множества 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:

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

При этом основу выделения множеств *объектов* и *ситуаций* составляет *синтаксический контекст существительного*. Указанный контекст для существительного, обозначающего некоторое понятие относительно заданной ситуации, есть представленная на *Плакате 3* последовательность из предикатного слова и соподчиненных друг другу существительных. Роль объекта относительно ситуации, обозначаемой предикатным словом, определяется типом отношения синтаксического подчинения между указанным словом и словом справа от него в последовательности. Тип синтаксического отношения характеризуется предлогом для связи главного слова с зависимым и падежом зависимого. Транзитивность отношения подчинения, которая следует из соотношения смыслов соподчиненных

слов, позволяет утверждать, что любое существительное рассматриваемой последовательности обозначает объект, значимый в заданной ситуации.

На основе выявленных соответствий «объект–ситуация–роль» выделяются группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях. Данная задача семантической кластеризации наиболее естественно решается методами Анализа Формальных Понятий (АФП, *Плакат 4*). При этом тернарному отношению между множествами текстов, объектов и ситуаций ставится в соответствие формальный контекст и строится решетка Формальных Понятий для исходного множества текстов. Анализ смысловой близости текстов таким образом сводится к исследованию качественных характеристик решетки. Визуализация решетки диаграмой линий позволяет графически отображать группировку текстов.

Тем не менее, *актуальной* является *проблема точности синтаксического анализа* как инструмента выделения понятий и их признаков. Известные синтаксические анализаторы, в частности, Cognitive Dwarf, реализуют стратегию разбора на основе наиболее вероятных связей. Вместе с тем, часто требуется исследовать природу выявляемых синтаксических связей. При неправильном разборе нужно установить причину использования той или иной стратегии (правила) с учетом особенности отражения ситуации, описываемой анализируемой фразой, в заданном ЕЯ.

Целью настоящей работы *является* разработка модели автоматического выделения и классификации наиболее вероятных синтаксических связей для множества СЭ-фраз.

Предлагаемое решение основано на закономерностях выражения смысла в заданном ЕЯ его носителем. При этом основополагающей является идея о разделении языкового опыта человека в соответствии с разделением концептуальной картины мира. Базовым здесь является понятие ситуации употребления ЕЯ как основы его генезиса.

Под *ситуацией употребления ЕЯ* понимают описание нового социального опыта (содержания совместных действий) средствами этого ЕЯ. Указанное описание выполняется в некоторой знаковой системе с целью обобщения и передачи знаний от человека к человеку. На *Плакате 5* представлена формальная модель языкового контекста, фиксируемого такой ситуацией, в виде совокупности трех составляющих: множества объектов, значимых в ситуации, множества отношений между объектами и множества форм языкового описания указанной ситуации.

Положим, что относительно заданного языкового контекста ситуация употребления ЕЯ задается множеством синонимичных фраз, каждая из которых описывает одну и ту же ситуацию действительности. При этом выбор той или иной фразы для описания ситуации является равновероятным. В силу произвольности отношений между значимыми в ситуации объектами предположим, что множество указанных отношений состоит из синтаксических отношений между словами, называющими эти объекты в рассматриваемом множестве синонимичных фраз. Тогда множество объектов следует рассматривать уже как множество понятий, значимых в заданной ситуации действительности, а само оно будет состоять из словесных обозначений этих понятий (включая понятие ситуации).

Таким образом, имеем следующую *задачу. Дано* множество синонимичных ЕЯ-фраз. *Требуется найти* отношения между словесными обозначениями понятий, значимых в описываемой ситуации действительности, используя указанные отношения в качестве признаков слов относительно заданной ситуации языкового употребления.

Рассмотрим текст как множество символов, которые его составляют. Тогда для любого текста из заданного синонимического множества справедливым будет выделить *(Плакат 6)* некоторую неизменную часть, которая является общей для всех текстов множества и изменяемую часть. На множестве символов последний выражаются синтагматические зависимости, которые задаются с помощью синтаксических отношений и определяют возможность сосуществования словоформ в линейном ряду.

Для установления синтаксического отношения значимой является *флексия* — та часть слова, которая изменяется при склонении (спряжении) и находится в его конце. На основе сочетания флексий выделяются морфологические зависимости. Поскольку указанные зависимости служат одним из способов *реализации синтаксического отношения*, то и само оно может быть выявлено попарным сравнением буквенного состава различных слов с выделением неизменной и флективной части.

Введем в рассмотрение *индексное множество* для неизменных частей всех слов, употребленных во всех фразах заданного синонимического множества. Назовем *моделью линейной структуры ЕЯ-фразы* последовательность индексов неизменных частей слов, присутствующих в этой фразе *(Плакат 7)*. При этом порядок индексов в модели идентичен порядку следования соответствующих слов во фразе, что позволяет однозначно восстановить фразу на множестве всех слов из всех фраз заданного синонимического множества. И наоборот, для любой из синонимичных фраз на заданном индексном множестве можно однозначно построить модель.

Для формирования искомого множества синтаксических отношений относительно заданной ситуации языкового употребления *необходимо найти* совокупность указанных моделей, удовлетворяющих *требованиям проективности. Модель линейной структуры ЕЯ-фразы* следует считать *проективной* в содержательном смысле, если все стрелки выявленных синтаксических связей могут быть проведены без пересечений по одну сторону прямой, на которой записана модель. Кроме того, если из позиции некоторого индекса выходят несколько стрелок, то эту позицию не должны накрывать стрелки, выходящие из позиций других индексов.

С учетом *линейной природы синтагм* дополним вышеуказанные требования следующим образом. Будем считать, что *модель* линейной структуры ЕЯ-фразы *проективна* относительно множества синтаксических отношений в заданной ситуации языкового употребления, если сумма длин всех связей относительно модели не превышает длины ее самой. При этом пара индексов, относительно которых задается связь, соответствует одной *синтагме. Связь* считается *допустимой для модели* линейной структуры ЕЯ-фразы, если в рассматриваемом синонимическом множестве существует пара фраз, модели линейных структур которых содержат в качестве подпоследовательности либо саму пару индексов, для которых определяется

связь, либо ее же, но записанную в обратном порядке.

Группировкой пар индексов, относительно которых заданы связи для моделей линейных структур, формируется *граф синтагм (Плакат 8)*, на основе которого строится *синтаксическое дерево-прецедент* заданного синонимического множества. Использованием маршрутов в данном дереве закономерности сосуществования слов в линейном ряду могут быть выявлены на основе *формального контекста сочетаемости флексий*, представленного на *Плакате 8*. При этом выделяемые классы синтаксических отношений соответствуют характеру изменения флективной части зависимосго слова. Поскольку исследуемая проблема точности синтаксического анализа характерна для ситуаций с двумя и более участниками, то и число дочерних узлов у корня дерева полагается больше одного.

Рассмотрим задачу поиска флексий для слов в составе расщепленных значений и конверсивов. Будем рассматривать Расщепленное Предикатное Значение (РПЗ) — совокупность вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию *(Плакат 9)*. Для РПЗ, как и для конверсивов (слов, обозначающих ситуацию с точки зрения разных ее участников) их неизменная часть не может быть найдена в буквенном составе слов всех фраз заданного синонимического множества. На *Плакате 10* представлены свойства модели линейной структуры ЕЯ-фразы, актуальные для поиска места *нераспознанного предикатного слова* в структуре *синтаксического дерева-прецедента*. При этом *основополагающим* для *доказательства* этих свойств является сделанное допущение о числе участников ситуации, обозначаемой предикатным словом, а также проективность модели. Для слов, которые вошли в РПЗ согласно *Лемме 2*, неизменная и флективная часть выделяются сравнением буквенного состава со всеми словами (не обязательно вошедшими в РПЗ) с нераспознанной неизменной частью в ЕЯ-фразах, не отвечающих условию *Теоремы 1*. Обязательным условием выделения флективной составляющей является преобладание сходств в буквенном составе сравниваемых слов *(Плакат 11)*. С учетом характерного для РПЗ направления ветвления в дереве подчинения, изначально выявленное дерево-прецедент преобразуется введением новых узлов для индексов слов в составе РПЗ согласно правилам, представленным на *Плакате 11*. В итоге основу формирования *формального контекста сочестаемости флексий* составляют те ЕЯ-фразы, которые наиболее полно описывают заданную ситуацию действительности относительно рассматриваемого языкового контекста.

Предложенная в работе модель процесса выделения и классификации синтаксических отношений была апробирована в ходе машинного эксперимента на материале результатов теста открытой формы *(Плакат 12)*. При этом основу построения *формального контекста сочетаемости флексий* составили максимально проективные ЕЯ-фразы с минимумом слов, не нашедших прообразы по буквенному составу *(Плакат 13)*. На *Плакате 14* представлена решетка Формальных Понятий *формального контекста сочестаемости флексий* для результирующего множества ЕЯ-фраз. При этом содержательная интерпретация решетки может

быть получена выделением морфологических классов слов на основе базиса импликаций *(Плакат 15)* для рассматриваемого формального контекста с учетом структуры последовательности соподчиненных слов в составе *синтаксического контекста существительного* согласно правилам, представленным на *Плакате 16*. Сами синтаксические отношения выделяются анализом наименьшей верхней грани каждой пары Формальных Понятий в решетке и образуют классы по сходству характера флексии зависимого слова. Отдельному классу соответствует область в решетке, а наименьшая верхняя грань множества Формальных Понятий этой области — прецеденту класса. В примере на *Плакате 14* классы отношений соответствуют словоизменению прилагательных (нежелательного, эмпирическ-ого) и существительных в составе генитивных конструкций (результат-ом переобучени-я, следстви-ем переобучени-я). Последний в силу транзитивности синтаксического отношения в рамках последовательности соподчиненных слов может включать сочетания существительного (вне генитивных конструкций) с глаголом.

Основу формирования решетки составляют те ЕЯ-фразы, которые максимально точно описывают ситуацию, а значит и более четко передают смысл. Следовательно, выявленные отношения будут соответствовать искомым наиболее вероятным синтаксическим связям относительно заданной ситуации языкового употребления.

Предложенная в работе *модель* позволяет решить *две* важные *задачи*, *актуальные для семантической кластеризации* ЕЯ-текстов.

Во-первых, автоматически *выделить лучший способ выражения нужной мысли в заданном ЕЯ*, что позволит избежать ошибок синтаксического анализа при использовании его как инструмента формирования объектов и признаков.

Во-вторых, *автоматизировать разработку синтаксических стратегий и правил* при исследовании случаев применения определенных грамматических конструкций в тематическом корпусе текстов. Качественные оценки формируемых знаний здесь могут быть даны на основе мер схожести решеток по аналогии с мерами схожести для Формальных Понятий.

Сферой рассмотрения настоящей работы были классы отношений для слов с изменяемой частью в конце словоформы. Тем не менее, чрезвычайно *интересным является дальнейшее развитие предложенного в работе метода* применительно к изменениям в составе основы слова. Здесь следует отметить беглые гласные, чередования гласных и согласных в составе основы, а также вариантные формы основ. В частности, отдельного рассмотрения заслуживает включение в рассматриваемые синтаксические контексты существительных имен числительных, для которых особенно актуально явление чередования в основах. Пример : «триста», «трехсот», «тремстам», «триста», «тремястами», «трехстах». В связи с этим другое *немаловажное направление дальнейших исследований* — распознавание слов-паронимов в составе синонимичных фраз. Наиболее плодотворные результаты данное исследование даст совместно с количественным изучением вариативности на уровне морфем и лексем русского языка.