Семантическая интерпретация в системах компьютерного анализа текста

Ермаков А.Е., Плешко В.В.

ООО "ЭР СИ О" (www.rco.ru)

Информационные технологии. - 2009. – N 6. – С. 2-7.

Аннотация

В статье описывается подход к построению семантического компонента в системах компьютерного анализа текста на естественном языке. Подход основан на применении специальных шаблонов к сети синтактико-семантических отношений между словами текста, которая строится синтаксическим анализатором. Шаблоны определяют способ интерпретации фрагментов сети в заданные фреймы, с идентификацией участников ситуаций и их ролей.

Ключевые слова: компьютерный анализ текста, семантическая интерпретация, семантическая сеть, синтаксический анализ, фреймы.

Введение

К числу современных компьютерных систем, использующих машинный анализ текста на естественном языке, традиционно относятся информационно-поисковые и вопросно-ответные системы, автоматические переводчики, а также широкий класс так называемых систем извлечения знания из текста, среди которых выделяют системы knowledge management (KM) и business intelligence (BI), системы сбора фактографической информации и ведения конкурентной разведки. Во всех подобных системах ключевым этапом автоматической обработки текста является семантическая интерпретация выражений естественного языка на основании определенной семантической модели мира (предметной области), результатом которой является формирование формальных структур, инвариантных к несодержательным лексико-грамматическим особенностям написания текста автором и соответствующих требованиям решаемой прагматической задачи.

Существуют различные подходы к построению семантического компонента систем анализа текста.

Первый подход, который следовало бы назвать "сильным", изначально возникший в рамках работ по машинному переводу, предполагает использование специальных семантических метаязыков для описания значения предложения [4]. Основоположниками этого подхода в нашей стране являются Апресян [1] и Мельчук [2], а среди их последователей сегодня особенно выделяется коллектив профессора Тузова, практическая работа которого подробно описана в книге [3]. В рамках этого подхода каждое значение

слова должно быть описано семантической формулой, а множество всех таких описаний представляет собой семантический словарь языка. Например, одно из значений слова *потушить* описывается формулой *Perf Caus(im,Fin Lab(vin,ОГОНЬ))*, где *огонь* - это базисное, далее не разложимое понятие, а *Perf, Caus, Fin, Lab* - это базисные семантические функции. Так, базисная функция *Lab(x,y)* обозначает, что аргумент, обозначаемый словом *x*, подвергается действию аргумента, обозначаемого словом *y*. Значение предложения описывается математическим выражением. Например, предложению *Пожарники потушили загоревшийся сарай* соответствует выражение *Perf Caus (ПОЖАРНИКИ,Fin Lab(Perf Incep Labo1(САРАЙ,ОГОНЬ),ОГОНЬ))*, которое в обратном переводе на русский язык имеет следующее значение: *Пожарники сделали так, что перестал подвергаться действию огня начавший подвергаться действию огня сарай*. Задача создания семантического метаязыка заключается в выборе системы таких базисных функций и понятий, которые позволяют описать значения всех других понятий и предложений, дальнейшее толкование которых либо невозможно, либо нецелесообразно. Так, метаязык, описанный в работе [3], содержит 72 базисные функции и около шестисот базисных понятий.

Будучи изначально создан для целей автоматического перевода текста с одного языка на другой, "сильный" подход к интерпретации текста на основе семантических метаязыков вынужденно претендует как на полноту охвата множества интерпретируемых выражений естественного языка, так и на точность (подробность) семантического описания этих выражений, что вызывает множество проблем при попытках практической реализации адекватного метаязыка [4], ни одна из которых по сей день не увенчалась успехом.

Прочие подходы, более "слабые", но вместе с тем более прагматичные, изначально подразумевает интерпретацию только тех фрагментов текста, которые описывают искомые отношения между сущностями предметной области или ситуации, в которые вовлечены эти сущности, не претендуя при этом на точность семантического описания. Например, приведенное выше предложение может быть проинтерпретировано просто как связь "*пожарник -> тушить пожар -> сарай*" или как ситуация "*пожар: объект=сарай*", в зависимости от назначения прикладной системы. К этой группе можно отнести самые разные методы семантической интерпретации, используемые в системах извлечения знаний из текста, вплоть до самых слабых, вообще не учитывающих синтаксис языка.

Статья посвящена методу семантической интерпретации, разработанному в компании "ЭР СИ О" (http://www.rco.ru). В соответствии с приведенной классификацией,

метод относится к группе "слабых", однако по факту полноты использования лингвистической информации всех уровней он является наиболее "сильным" в этой группе, во всяком случае, подобных методов автору не известно ни в России, ни за рубежом. Описываемый семантический компонент определяет, каким образом будут интерпретированы те или иные языковые конструкции, описывающие те или иные ситуации, в заданные фреймы, с идентификацией участников ситуации и их ролей. С этой целью для каждого типа ситуаций создаются особые синтактико-семантические шаблоны, позволяющие распознать и проинтерпретировать допустимые способы описания ситуации в тексте. Такие шаблоны применяются не к тексту, а к сети синтактико-семантических отношений между словами, которая строится синтаксическим анализатором, что обеспечивает высокую инвариантность шаблонов к особенностям поверхностно-синтаксической организации предложений и, как следствие, ускоряет разработку лингво-семантического описания предметной области.

Сеть синтактико-семантических отношений

В результате синтаксического анализа предложения и последующих трансформаций дерева синтаксических зависимостей между словами [5] формируется сеть синтактико-семантических отношений - семантическая сеть. Семантическая сеть содержит все сущности, упоминавшиеся в тексте предложения - наименования предметов и лиц, действий и признаков, связанные различными типами синтактико-семантических связей. Направление связи обычно соответствует направлению синтаксического подчинения слов. Пример семантической сети представлен на рисунке 1.

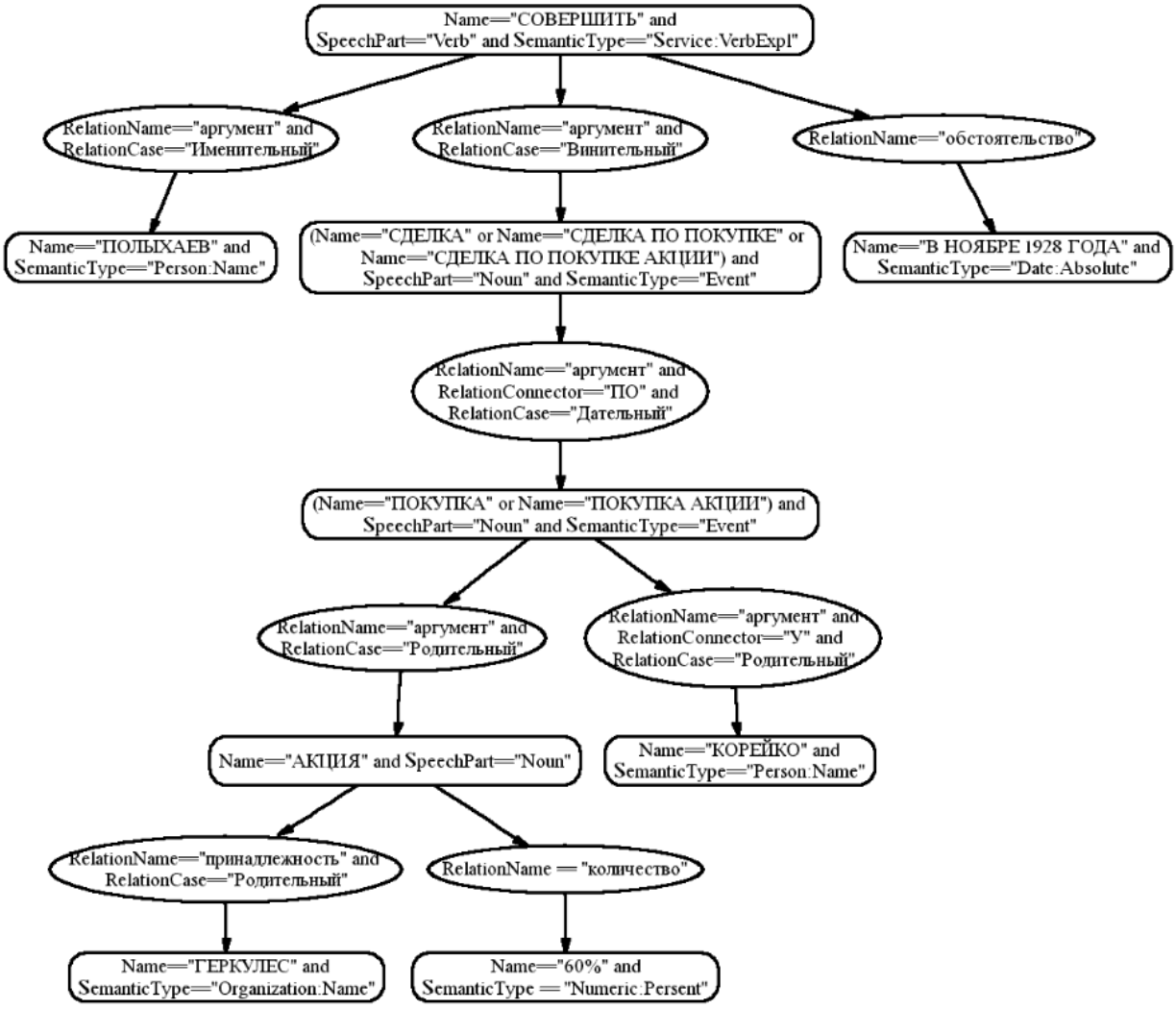

Рисунок 1. Пример семантической сети, соответствующей предложению "*В ноябре 1928 года Полыхаев совершил сделку по покупке 60% акций ООО “Геркулес” у Корейко*"

Узлы и связи в сети имеют набор следующих основных атрибутов:

- часть речи слова, соответствующего узлу (*SpeechPart*).
- семантический разряд референта узла (*SemanticType*), например: *персона*, *организация*, *географическое место*, *действие/состояние*, *предмет* и др.
- строка текста, соответствующего узлу, в нормальной форме (*Name*). Для именных групп может иметь несколько значений, которые представляют все цельные словосочетания, образованные от главного существительного в узле, например: *сделка по покупке акции*, *сделка по покупке*, *сделка*.
- тип синтактико-семантической связи между узлами (*RelationType*), например: *аргумент (совершить -> сделку)*, *принадлежность (акция -> Геркулеса)*, *обстоятельство (совершить -> в 1928)*.

- семантический падеж (*RelationCase*) и коннектор (*RelationConnector*) – предлог или союз, при помощи которого устанавливается связь. Комбинации условий *RelationConnector* + *RelationCase* представляют альтернативу традиционным семантическим ролям (*субъект*, *объект*, *инструмент*, *локатив* и т.п.), инвентарь которых в лингвистике так и не определен окончательно. При этом формально различные грамматические падежи слов в тексте, выбор которых определяется поверхностно-синтаксической организацией фразы, отображаются в один и тот же семантический падеж. Например, семантический именительный субъекта действия и винительный объекта действия соответствуют одноименным грамматическим падежам в активном залоге (*программист написал программу*), в пассивном выражаются грамматическим творительным и именительным соответственно (*программистом написана программа*), а в причастном обороте вообще могут выражаться любыми грамматическими падежами (*программисту, написавшему...; о программе, написанной* ...).

Семантическая сеть инвариантна к синтаксической структуре и порядку слов с точностью до структуры пропозиции, выбранной автором для описания ситуации. Например, конструкциям "*Корейко купил акции Геркулеса у Берлаги*" и "*акциями Геркулеса, купленными Корейко у Берлаги*" будут соответствовать одинаковые сети. В то же время пропозициям вида "*Корейко становится покупателем акций Геркулеса*" и "*покупка акций Геркулеса – дело рук Корейко*" будут соответствовать иные сети. Описанная семантическая сеть является промежуточным уровнем представления между собственно семантической схемой ситуации и ее конкретным языковым описанием, т.е., представлением глубинно-синтаксического уровня, абстрагированным от особенностей поверхностного синтаксиса.

Фреймы и семантические шаблоны

Логическая схема ситуации в терминологии искусственного интеллекта называется фреймом. Фрейм имеет имя, которое идентифицирует тип описываемых им ситуаций (например, *купля-продажа акций*), а также содержит слоты, которые имеют свои имена, идентифицирующие роли участников ситуации. Для конкретной ситуации часть слотов может быть заполнены именами ее конкретных участников, упомянутых в тексте (*покупатель=Корейко, эмитент акций=Геркулес, количество акций=60%, дата=1928, продавец=?, сумма сделки=?*).

Для семантической интерпретации каждого способа описания ситуации в тексте используется соответствующий синтактико-семантический шаблон. На рисунке 2 приведен пример такого шаблона, соответствующего пропозиции вида *Покупатель совершает действие по приобретению акций предприятия-эмитента у продавца*.

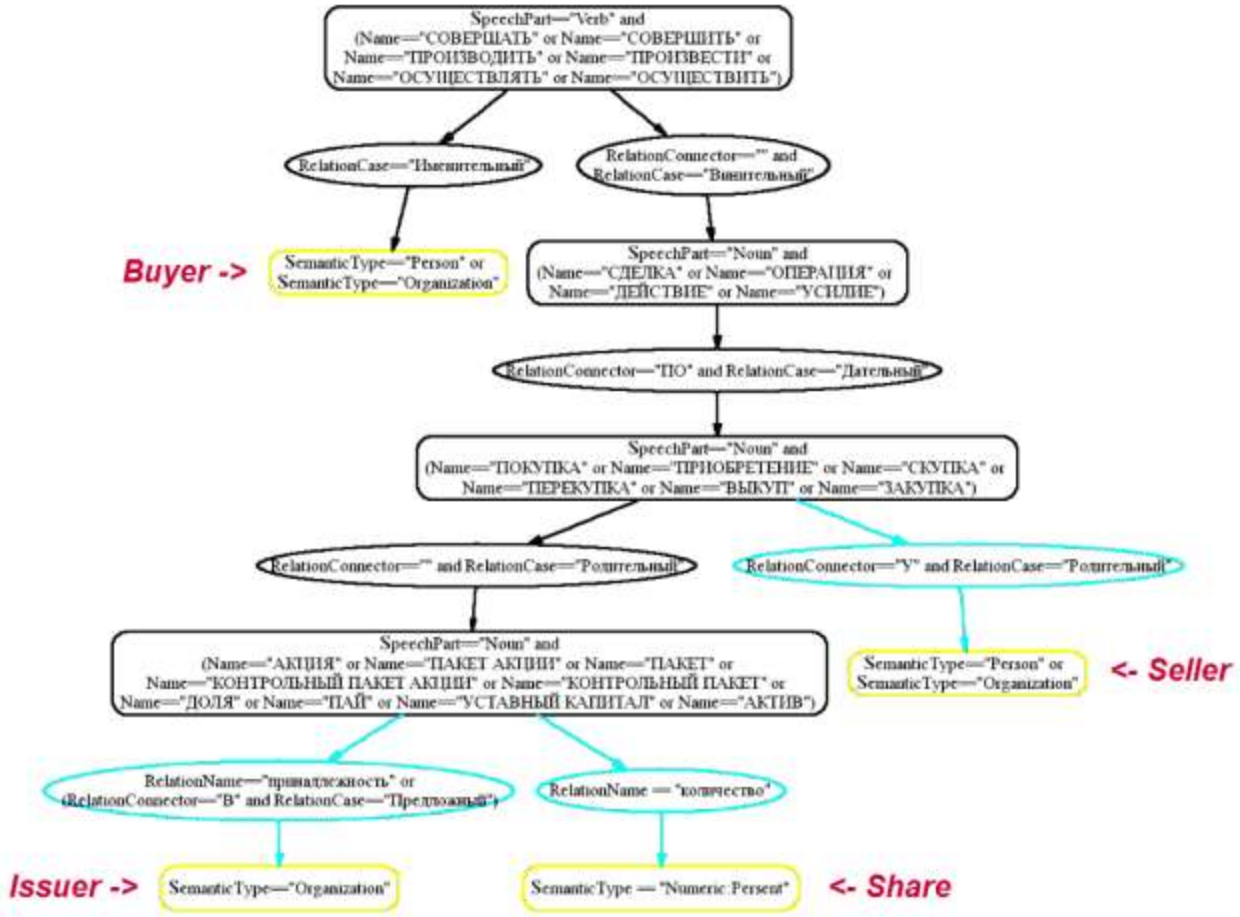

Рисунок 2. Пример синтактико-семантического шаблона, распознающего ситуации, выраженные пропозицией вида "*Покупатель совершает действие по приобретению акций предприятия-эмитента у продавца*".

Как видно, синтактико-семантический шаблон задается в виде сети, изоморфной искомой в тексте, в узлах и связях которой при помощи логических выражений указываются условия, которым должны удовлетворять узлы и связи искомой сети. Обычно в некоторых узлах шаблона содержатся конкретные слова, которые должны присутствовать в тексте. Другие узлы - валентности шаблона - соответствуют искомым участникам и дополнительно содержат обозначения их ролей - это имена слотов фрейма, заполняемых словами из текста при нахождении фрагмента семантической сети, соответствующего шаблону. Так, на рисунке 2 узлы с именами *Buyer*, *Issuer*, *Seller* и *Share* представляют возможных участников ситуации "*покупка акций*" в ролях "*Покупатель*", "*Эмитент акций*", "*Продавец*", "*Размер доли*" соответственно. Светлые связи к

фигурантам *Seller*, *Issuer* и *Share* помечены как факультативные, так как соответствующие участники могут не упоминаться в тексте.

В рамках описываемой модели, семантическая интерпретация описания ситуации в тексте есть поиск в его семантической сети такой подсети, которая изоморфна шаблону, с заполнением слотов соответствующего фрейма именами участников ситуации из текста в соответствии с ролями, указанными в узлах шаблона.

На практике возможны такие случаи, когда синтаксический анализатор не может установить связь между словами, опираясь на заложенные в него общие правила грамматики, например, в текстах особого стиля: *Соучредитель ООО "Геркулес" (20 %) – ЗАО "Рога и копыта"*. Чтобы решить такие проблемы, в семантическую сеть добавляются связи особого типа (*RelationType == "next"*), которые связывают в цепочку идущие друг за другом в предложении слова и знаки препинания, причем "перепрыгивая" через синтаксически подчиненные слова в именных группах и связывая только их вершины, что позволяет писать шаблоны, инвариантные к количеству слов в словосочетаниях. В итоге, совокупность узлов сети всегда представляет собой полносвязный граф, что позволяет описывать на основании единого формализма как универсальные общеязыковые, так и специфические для предметной области и стиля текста типы текстовых конструкций.

Расширенная семантическая интерпретация

Для уменьшения количества шаблонов, описывающих целевые фреймы, существует возможность создавать служебные шаблоны, назначение которых состоит не в заполнении фреймов, а в добавлении к семантической сети текста дополнительных узлов и связей с заданными атрибутами. В ходе семантической интерпретации все шаблоны применяются в определенном порядке, и каждый следующий шаблон обрабатывает сеть, которая являются совместным результатом работы синтаксического анализатора текста и всех предыдущих шаблонов, содержащих порождаемые узлы и связи. Так, если для всех типов извлекаемых фреймов все конструкции, построенные по типу "*Субъект принимает решение действовать (о действии)* ..." предполагают такую же интерпретацию, как и конструкции типа "*Субъект действует*", то, вместо добавления соответствующего шаблона для каждого типа фрейма, целесообразно написать один служебный шаблон (Рис. 3), которые будут добавлять в сеть связь от названия действия к субъекту, вследствие чего подобные описания ситуаций будут распознаваться по простым шаблонам типа "*Субъект действует*".

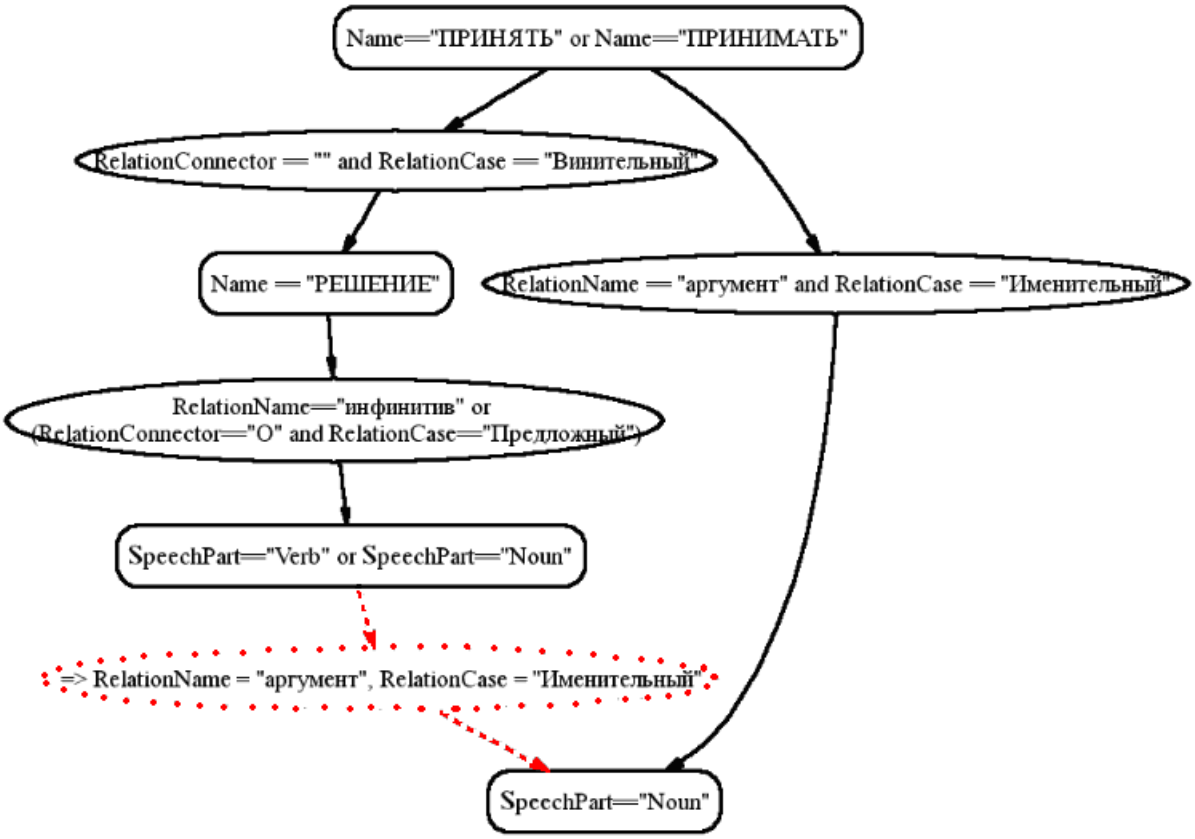

Рис. 3. Пример служебного шаблона, распознающего фрагменты семантической сети, соответствующие конструкциям вида "*Субъект принимает решение действовать (о действии)*". При распознавании в сеть добавляется новая связь (обозначена пунктиром) между предикатом действия, выраженным глаголом или существительным, и субъектом действия, выраженным существительным.

Кроме того, для уменьшения количества создаваемых шаблонов компонент семантической интерпретации поддерживает средства их параметризации, которые позволяют подключать внешние словари к требуемым ролям шаблонов, указывая слова и семантические разряды допустимых или, наоборот, недопустимых слов, что позволяет выделять фреймы с одинаковой структурой слотов, но разными типами, на основании одного и того же множества общих шаблонов. Так, например, имея множество общих шаблонов, описывающих ситуацию "*приобретение-покупка чего-либо*", можно посредством подключения словарей к роли "*Товар*" определить фреймы типа "*приобретение предприятий*", "*приобретение акций*", "*приобретение недвижимости*", а также "*приобретение прочих вещей*", исключив из числа *прочих вещей* все *предприятия*, *акции*, *недвижимость*, а также *влияние*, *доверие* и им подобные непредметные сущности.

Ключевой особенностью текстов некоторых стилей (биографии, описания товаров) является высокая плотность таких связей между словами, которые не выражаются грамматическими средствами - анафорических связей. Большинство предложений в подобных текстах либо бессубъектно (*Родился в 1958 году. Работает директором ООО "Геркулес"*), либо номинативно (*1958 года рождения. Директор ООО "Геркулес"*), либо вообще разорвано в списках (*Является владельцем акций следующих предприятий: - ООО "Геркулес", основано в 1928 году .... – ООО "Рога и копыта", основано в 1924 году ... - ООО "Ударник", основано в 1925...*). Для поиска тех факультативных участников ситуации, которые не были найдены в предложении в результате применения шаблона, разработан специальный механизм поиска анафорических связей по всему тексту, опирающийся на специальные лингвистические правила и соответствующую маркировку узлов шаблонов [6].

Наконец, настройки семантического интерпретатора позволяют фильтровать описания тех ситуаций, которые соответствуют реальным событиям или фактам, на основании наличия в тексте общеязыковых показателей отрицания и нереальности (*не, якобы, если* и т.п.). В итоге, из числа найденных могут быть исключены те фреймы, которые соответствуют не реальным ситуациям (*Корейко купил бы акции "Геркулеса"*), и те участники, которые реального участия в ситуации не принимали (*Корейко купил акции не "Геркулеса"*).

Настройка семантических шаблонов

Семантические шаблоны задаются на формальном языке описания графов DOT (http://www.graphviz.org). Для удобства их разработки используется приложение с графическим интерфейсом, которое строит сеть на основе типовой фразы естественного языка, т.е. обучает семантический компонент на примерах. В итоге, граф шаблона строится автоматически, после чего человеку остается проставить требуемые логические выражения в узлы (обычно требуется добавление синонимов), указать роли искомых участников ситуации, пометить обязательных и факультативных участников. Окончательно шаблон сохраняется в формализме DOT, готовый для загрузки компонентом семантической интерпретации, который обеспечивает поиск изоморфизмов и заполнение фреймов по семантической сети при анализе текста.

Для автоматизированного построения большого количества семантических шаблонов используется следующий метод их пакетного построения:

1. Языковая основа каждого шаблона описывается типовой конструкцией на естественном языке с соблюдением следующих соглашений:

- типовая конструкция состоит из известных слов естественного языка и представляет собой грамматически правильную фразу. Каждое слово конструкции может представлять собой макроопределение - отсылать к одноименному классу слов, способных занимать это слово-место. При этом синтаксическая форма конструкции однозначно определяет типы связей, устанавливаемых между соответствующими узлами в графе шаблона;

- макро-определение может быть описано непосредственно в составе конструкции или вынесено в отдельную запись и имеет одну из четырех допустимых форм: *Макроопределение=слово1=слово2=...=словоN* или *Макроопределение={ слово1, слово2, словоN }* или *Макроопределение=[логическое выражение, допустимое в графе шаблона]* или *Макроопределение:РольУчастника*. Первые три формы транслируются в логическое выражение и определяют ограничения на узел шаблона. Последняя форма транслируется в роль участника ситуации, с возможным маркером его факультативности "~", например: *Продавец:~Seller*. Во избежание возможных конфликтов, вызванных многозначностью слов, макросы могут иметь глобальную и локальную область видимости.

Пример двух типовых конструкций с макросами, описывающих два способа выражения ситуации покупки акций:

```
Покупатель:Buyer; Покупатель = { SemanticType == "Person" or SemanticType == "Organization" };
Продавец:~Seller; Продавец = { SemanticType == "Person" or SemanticType == "Organization" };
Эмитент:~Issuer; Эмитент = { SemanticType == "Organization" };
Покупка = { приобретение, скупка, перекупка, выкуп, закупка };
Покупать = { купить, приобретать, приобрести, скупать, скупить … };
Акции = { пакет, пакет акций, контрольный пакет, контрольный пакет акций, доля, пай, уставный капитал, актив }
Покупатель Покупает Акции Эмитетна у Продавца;
Покупатель совершает=производить=произвести=осуществлять=осуществить сделку=операция=действие=усилие по Покупке Акций Эмитетна у Продавца;
```

2. Пакет типовых конструкций обрабатывается программой-транслятором, которая подвергает синтаксическому анализу каждую конструкцию, строя соответствующую ей семантическую сеть – граф шаблона, и подставляя вместо слов – обозначений макросов – соответствующие им определения. В ходе трансляции создается журнал, в котором фиксируются следующие основные ошибки:

- ошибки разбора текста описания шаблонов, связанные с нарушением формата.

- ошибки или неоднозначности разбора типовой конструкции: полный синтаксический разбор не удался или допустимо несколько вариантов разбора.

- слово в типовой конструкции, написанное с большой буквы, не удалось отождествить с макросом. Возможно, соответствующий макрос не определен или нормальная форма слова не совпадает с макросом.

3. После анализа сообщений об ошибках эксперт-настройщик устраняет их одним из двух возможных способов:

- корректирует текст описания: исправляет синтаксические ошибки, добавляет определения макросов, переформулирует текст омонимичных конструкций, после чего вновь запускает программу-транслятор;

- корректирует непосредственно семантические шаблоны, используя графическое приложение.

Заключение

Описанный подход к семантической интерпретации внедрен в компонент лингвистического анализа текста RCO Fact Extractor и успешно апробирован на базе русского и английского языков, обеспечивая в среднем около 95% точности и 60% полноты при извлечении из текста описаний событий и фактов в соответствии с заданными семантическими шаблонами. В настоящий момент для русского языка уже разработано более 600 шаблонов, которые покрывают более 70 типов ситуаций, связанных с экономической и общественно-политической деятельностью персон и организаций. Структура фреймов (типы ситуаций, состав ролей участников) и соответствующие шаблоны разрабатывались "с нуля" на основе анализа текстовых примеров, предоставляемых заказчиками. Для английского языка, работа с которым недавно начата, в качестве полезного ресурса предполагается использовать базу данных проекта FrameNet (http://framenet.icsi.berkeley.edu/), которая содержит разработанные структуры более чем 800 фреймов, с примерами описаний соответствующих ситуаций более чем в 130 тысячах типовых предложений.

Литература

1. Апресян Ю.Д. Лексическая семантика. М.: Наука, 1974, 366с.

2. Мельчук, И.А. Опыт теории лингвистических моделей «Смысл↔Текст». М.: Наука, 1974.

3. Тузов В.А. Компьютерная лингвистика. Опыт построения компьютерных словарей. - СПб.: Изд-во СПбГУ, 2002,650 с.

4. Кобозева И.М. Лингвистическая семантика. М.: Эдиториал УРСС, 2000. – 352 с.

5. Ермаков А.Е. Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза. // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003. – Москва, Наука, 2003. - С. 136-140.

6. Ермаков А.Е. Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2007. – Москва, Наука, 2007. - С.131-135.