Поиск фактов в тексте естественного языка на основе сетевых описаний

Киселев С.Л., ИИК “Белый ветер”
Ермаков А.Е., Плешко В.В., ООО “Гарант-Парк-Интернет”

Доклад посвящен технологии автоматического анализа текста русского языка и поиска в нем описания фактов заданного типа, в том числе извлечения требуемых фигурантов факта и сопутствующих обстоятельств. Описывается представление текста в форме сети синтактико-семантических отношений, которая инвариантна к форме описания фактов с точностью до выбранной автором структуры пропозиции. Для поиска фактов используются шаблоны в форме сетей с заданными ограничениями на атрибуты узлов и связей, которые позволяют находить, преобразовывать и интерпретировать требуемые семантические структуры в сети текста.

Речь пойдет о технологии, которая позволяет найти в тексте описания фактов заданного типа, например, “поездки” или “поддержка на выборах”, и извлечь требуемую информацию, связанную с фактами - имена задействованных участников, обстоятельства места и времени и другое. Основная сфера приложения технологии - это аналитические задачи из области компьютерной разведки, требующие высокоточного отбора информации по заданным смысловым критериям, например, автоматизированное составление досье на целевые персоны или организации.

Реализованная нами технология фактографического поиска опирается на модель содержания текста в форме семантической сети. Семантическая сеть содержит все полнозначные слова и словосочетания, упоминавшиеся в тексте - наименования объектов, действий и признаков, связанные различными типами синтактико-семантических связей.

Элементарная сеть представляет результат синтаксического анализа и постсинтаксических трансформаций дерева синтаксических зависимостей между словами в отдельном предложении. Некоторые принципы используемого нами синтаксического анализа были описаны в [1,2], а полная информация о синтаксическом анализаторе RCO Syntactic Engine представлена на сайте http://www.rco.ru. Полная сеть текста есть результат объединения отдельных семантических сетей на основе узлов, соответствующих кореферентным именам объектов.

Узлы и связи в сети имеют набор следующих атрибутов:

- Name - строка текста, соответствующая узлу. Может иметь несколько значений, каждое из которых соответствует цельному словосочетанию, образованному от ключевого существительного в узле, например: “*новый указ президента*”, “*указ президента*”, “*указ*”, или одному из кореферентных имен объекта в тексте “*Василий Иванов*”, “*директор*”, “*известный предприниматель*”.
- SemanticCategory – семантический разряд ключевого слова, соответствующего узлу.
- RelationType – тип синтактико-семантической связи между узлами, например “*аргумент*”, “*признак*”, “*принадлежность*”.
- RelationRole – семантическая роль, определенная для связей предиката с аргументом, получаемая обычно из словаря моделей управления, например “*субъект*”, “*объект*”, “*инструмент*”.
- RelationCase, RelationConnector – семантический падеж и коннектор (предлог, союз), при помощи которых устанавливается связь предиката с аргументом. Представляют альтернативу семантической

роли, так как роль не всегда может быть установлена. Один и тот же семантический падеж может соответствовать различным грамматическими падежами в зависимости от построения фразы. Например, семантический именительный субъекта действия и винительный объекта соответствуют одноименным грамматическим падежам в активном залоге, а в пассивном выражаются грамматическим творительным и именительным соответственно.

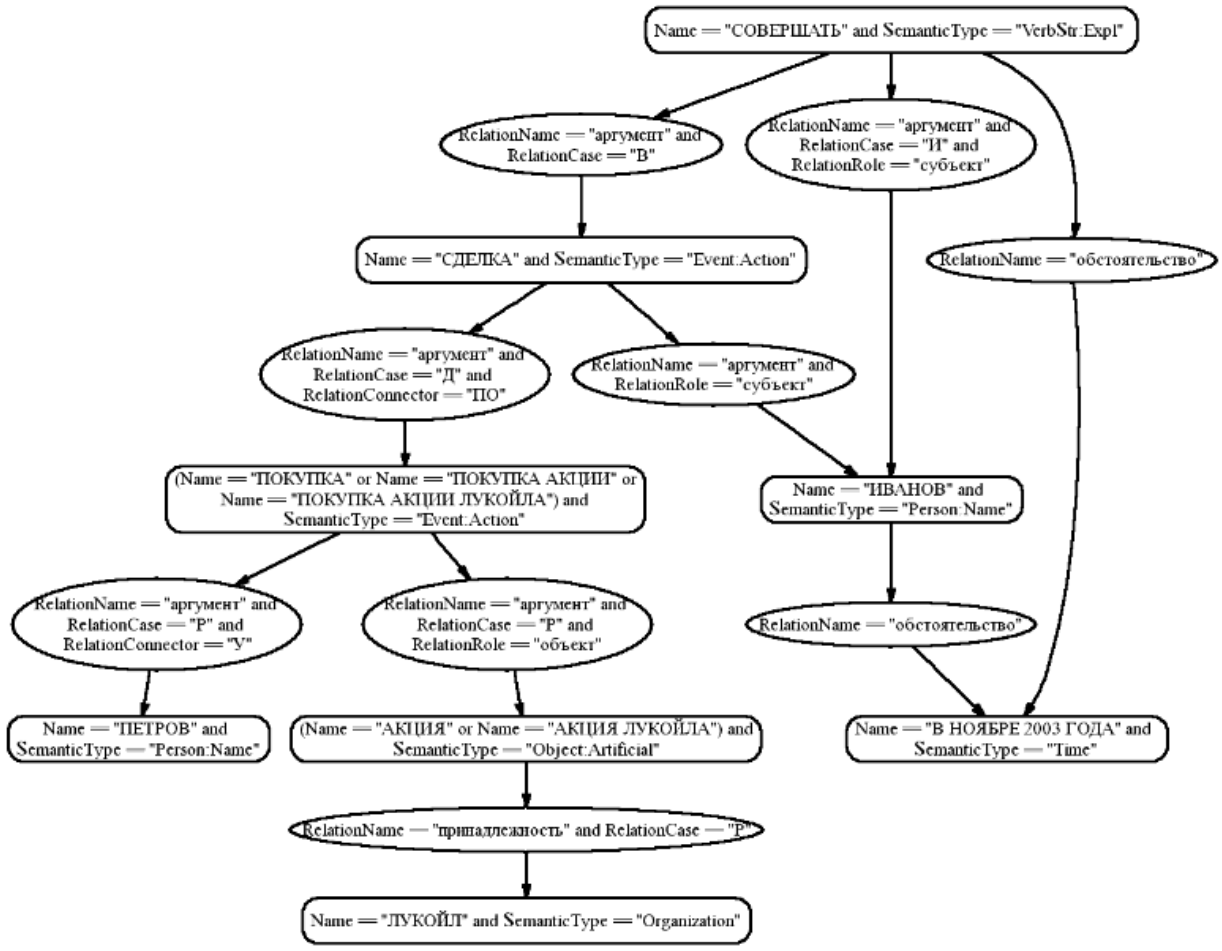

Рис. 1. Пример семантической сети, соответствующей предложению "*В ноябре 2003 года Ивановым была совершена сделка по покупке акций Лукойла у Петрова*".

Представление содержания текста в форме семантической сети позволяет абстрагироваться от многих особенностей его коммуникативной организации. Такая сеть инвариантна к синтаксической структуре предложений и порядку слов с точностью до структуры пропозиции, выбранной автором для описания ситуации. Например, конструкциям "*Иванов купил акции*" и "*акциях, купленных Ивановым*" будут соответствовать одинаковые сети. В то же время пропозициям вида "*Иванов становится покупателем акций Лукойла*" и "*покупка акций Лукойла – дело рук Иванова*" будут соответствовать иные сети. Вследствие этого семантическая сеть является промежуточным уровнем представления между собственно семантической схемой ситуации и ее языковым описанием.

Модель факта задается множеством лингвистических описаний (ЛО), каждое из которых описывает множество изоморфных семантических сетей, соответствующих некоторому типовому способу описания факта в тексте. Основными элементами ЛО являются:

- участники ситуации – узлы сети, которые соответствуют текстовым единицам, извлекаемым в качестве значений фигурантов факта. Например, в ситуации покупки акций потенциально присутствуют участники с ролями "*продавец*", "*покупатель*" и "*эмитент акций*", а "*товаром*" всегда являются акции.

- вспомогательные элементы - узлы сети с заданными ограничениями на атрибуты, которые позволяют распознать присутствие описания факта в тексте. Обычно они соответствуют наименованию ситуации (“*покупка*”, “*покупать*”, “*приобретать*”) или именам обязательных участников, более точно идентифицирующим ее (тип товара: “*акция*”, “*контрольный пакет*”).

- схема ситуации – набор связей между участниками и вспомогательными элементами с заданными ограничениями на атрибуты связей. Схема ситуации соответствует связям в семантической сети простого неосложненного предложения, свободного от дополнительных участников, обстоятельств, определений и прочего.

Поиск факта есть поиск в семантической сети текста такой подсети, которая изоморфна одному из ЛО. Если подсеть найдена, факт считается установленным, после чего производится извлечение текстовых значений фигурантов факта (атрибут Name) и их интерпретация в соответствии с ролями, заданными в соответствующих узлах ЛО.

Дополнительно в схеме ЛО могут присутствовать необязательные узлы, которые соответствуют дополнительным участникам или обстоятельствам. После нахождения изоморфизма производится поиск необязательных узлов, и текстовые значения соответствующих фигурантов также извлекаются.

Существует возможность вводить в ЛО порождаемые объекты и связи с любыми заданными атрибутами, которые добавляются к заданным узлами сети текста при нахождении изоморфизма. В ходе работы все ЛО применяются в заданном порядке, и каждое следующее ЛО может обрабатывать подсети, которые являются совместным результатом работы анализатора текста и всех предыдущих ЛО. В результате по мере срабатывания ЛО сеть синтактико-семантических отношений может быть постепенно превращена в сеть смысловых отношений, сохраняя при этом все исходные связи, явно выраженные в тексте.

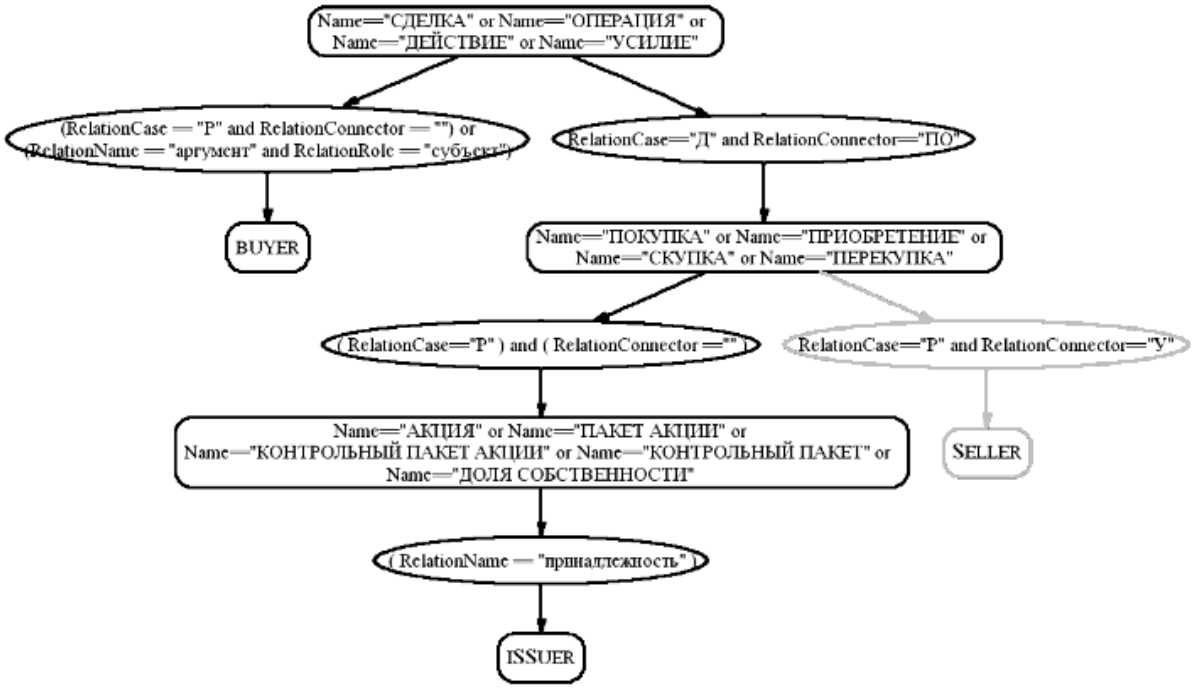

Рис. 2. Пример ЛО, покрывающего множество описаний факта в форме пропозиции вида “*Покупатель совершает действие по приобретению у продавца акций предприятия*”.

В ЛО на рисунке 2 три узла, обозначенные метками BUYER, ISSUER и SELLER, представляют возможных фигурантов факта “покупка акций” – покупатель, эмитент и продавец соответственно. Узел SELLER вместе с идущей к нему связью является необязательным, так как продавец может и не указываться в тексте, и именно пара “покупатель-эмитент” представляет интерес для факта покупки акций.

ЛО задаются на формальном языке описания графов, который позволяет определить структуру сети и наложить ограничения на атрибуты узлов и связей в виде логических выражений. Для удобства настройки

ЛО используется модуль с графическим интерфейсом, позволяющий построить сеть на основе типовой фразы русского языка. После добавления требуемых ограничений на узлы сети, указания обязательных, необязательных, порожденных элементов и ролей участников факта, ЛО сохраняется в нужном формализме, готовое для работы системы фактографического поиска.

В практических задачах, требующих высокой полноты поиска, мощность класса ЛО, необходимых для выделения факта одного типа, может колебаться от десятков (например, для фактов "заключение договоров") до сотен (например, для фактов "конфликты"). Формирование множества таких классов, описывающих предметную область, относится к сфере инженерии знаний и требует операций декомпозиции области на элементарные факты, их классификации и упорядочивания. Качественное описание иерархии фактов требует формирования многоуровневой гетерархической структуры, так как один и то же факт на любом из уровней может допускать несколько целевых интерпретаций. Способы решения данной задачи выходят за рамки задач компьютерной лингвистики и относятся сфере приложения методов искусственного интеллекта.

В настоящее время нами разработаны ЛО для извлечения из текста нескольких десятков типов фактов, которые связаны с действиями юридических лиц и VIP-персон, часто освещаемыми в СМИ.

Литература

1. Ермаков А.Е. Неполный синтаксический анализ текста в информационно-поисковых системах. // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2002. В двух томах. Т.2. "Прикладные проблемы". – Москва, Наука, 2002. - С. 180-185.
2. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста. // Информационные технологии. - 2002. – N 7. – С. 30-34.