Об одном подходе к обработке естественно-языковых данных на основе анализа семантических сетей

В. Дерецкий
Институт программных систем НАН Украины
252187 г. Киев, просп.Академика Глушкова, 40, Институт программных систем НАН Украины
тел.: +380 (44) 266 43 42
E-mail: dva@isofts.kiev.ua

This paper presents a natural language processing system implementations on a memory based parsing approach, the Semantic Network Array Processing. Our system used the approach, which linguistic information is stored as phrasal patterns in a semantic network knowledge base distributed over the memory of the computer. Parsing is performed by recognizing and linking phrasal patterns that reflect a sentence interpretation. We have developed a system CONTENT capable of processing newswire articles from a particular domain.

Index terms: Natural languages processing. Semantic network. Information extraction. Information retrieval.

1 Введение

Эффективное использование знаний, содержащихся в текстах, требует новых стратегий обработки информации, которые бы учитывали семантические законы естественного языка. Семантический аспект важен для всех направлений информационной индустрии таких, например, как сбор, накопление, хранение, доступ и интерпретация полученной информации.

Методы автоматизированной обработки текста, представленного на естественном языке исследуются в рамках предмета инженерной лингвистики. Основная проблема инженерной лингвистики может быть выражена таким образом: можно ли разработать алгоритмы, которые могли бы позволить обрабатывать естественно-языковые данные так же, как это делает специалист (человек). Решение этой проблемы связано с раскрытием таких понятий как: что такое значение текста, извлечение значения из текста, представление значения текста, всегда ли нужно в полной мере определять значение при решении конкретной задачи "понимания".

Проблема лингвистики и философии языка - что такое значение текста в данной работе сведена к его записи на специально - разработанном языке-посреднике, основанном на использовании семантических сетей.

Настоящая работа направлена на разработку методов извлечения информации из текста. Цель исследований данного направления [3] заключается в построении систем, которые находят в тексте и связывают релевантную для пользователей информацию, игнорируя избыточную и нерелевантную информацию. Базовыми компонентами типичной системы извлечения текстовой информации являются: фильтрация, морфологический анализ, синтаксический анализ, дискурсная реферация и синтез результата.

Системы извлечения информации могут иметь широкое применения. Например, обрабатывая новости, можно выдавать только интересующую конкретного пользователя коммерческую или финансовую информацию. Поиск новых публикаций в конкретной области исследований, определение тенденций развития интересующих вопросов, отраженных в текстах и другие.

В рамках настоящей работы, разрабатывается подход и на его основе программно-инструментальный комплекс КОНТЕНТ [10], направленные на "понимание"и извлечение информации из текста. В отличие от приведенной выше классической схемы извлечения информации в данной работе применяется интегральный подход [9, 1, 6].

КОНТЕНТ - это экспериментальная система, способная "понимать"тексты и обобщать фактический материал, который был "понят"при "чтении"текста. Работа системы основана на использовании ряда источников знаний, основными из которых являются знания о предметной области "понимаемого"явления, семантические знания о словах и выражениях, составляющих текст и др. Настоящую версию системы можно отнести к типу "поверхностного понимания"текста, при котором из текста извлекается лишь его основное содержание (факты, события) и игнорируются глубинные особенности содержания текста. Хотя следует отметить, что метод, положенный в основу системы, позволяет последовательно увеличивать глубину понимания текста.

В результате "поверхностного понимания"система готовит обобщающий аналитический материал в виде селективных текстов, списков семантических категорий, таблиц и графиков, содержащих статистические данные, обобщающие "прочитанный"материал.

2 Модель понимания текста

Модель "понимания"текста в системе КОНТЕНТ основана на методе концептуального анализа семантических фреймов. Его роль в процессе "понимания"состоит в том, чтобы построить по возможности наиболее полное значение высказывания в терминах заданных концептуальных зависимостей, основываясь на значении слов и выражений, встретившихся в данном высказывании.

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

Базовая модель "понимания"текста в системе КОНТЕНТ сводится к следующему (см. рис.1):

Описываются знания о предметной области в виде множества фреймов. Каждый фрейм представляется в виде системы концептов и концептуальных зависимостей. В данной работе приведен пример фрейма "Сотрудничество между Украиной и НАТО".

Языковые знания о словах представляются в словаре. Определение слова или словоформы состоит из его значения и синтаксической информации. Значение определяется концептуальной зависимостью, из совокупности знаний о предметной области. Синтаксические знания определяются набором синтаксических признаков, ассоциируемых с данным словом или словоформой.

Анализ входного высказывания (определение значения высказывания) состоит в пословном соотнесении текста и модели предметной области. Осуществляется поиск в словаре, содержащем языковые знания. В соответствии с найденным значением слова или словоформы инициируется агент фрейма. Инициируются так же агенты фрейма, которые используют синтаксическую информацию о словоформе и уточняют значение высказывания или предложения. В результате анализа всего предложения построенная цепочка концептов сохраняется в памяти базы знаний. Осуществляется анализ памяти базы знаний с целью формирования смысловых конструкций. Формирование смысловых конструкций состоит в создании новых фреймов событий или ситуаций, релевантных контексту.

Далее рассмотрим основные компоненты и их взаимодействие в предложенной модели.

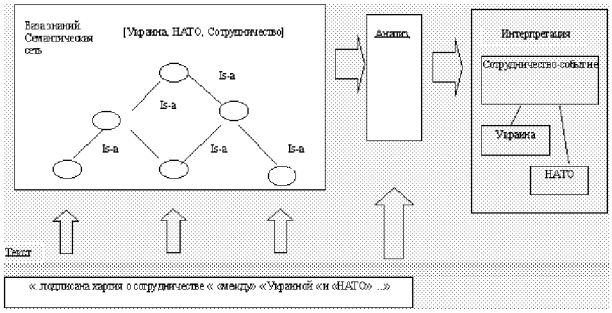

Рис. 1: Модель "понимания"текста в системе КОНТЕНТ

3 Представление знаний о предметной области

Знания системы КОНТЕНТ о предметной области - это множество семантических примитивов, представляющих данную предметную область в формате концептуальной зависимости. Каждый семантический примитив - концепт, который состоит из заголовка (header), за которым следует некоторый набор поименованных позиций - слотов. Ядро и поименованные слоты вместе составляют фрейм понятия. Разные понятия могут ссылаться между собой, а именно, фрейм одного понятия может быть помещен в слот другого понятия в соответствии с определенными ограничениями. Возможности сочетания одного понятия с другим в предметной области реального мира потребует достаточно сложных правил рассуждения. В системе КОНТЕНТ используется следующий подход - в результате любого процесса сложного рассуждения получается суждение, аналогичное по своей структуре конструкции фрейма.

Идея представления знаний в виде сети концептов и устанавливаемых взаимосвязей посредством программных маркеров, является новым подходом к обработке естественно-языковых данных (memory-based parsing). Подход состоит в реализации правил разметки концептуальной сети или грамматик над большим объемом памяти базы знаний. Анализ (parsing) представляется как формирование и хранение маркеров разметки входных предложений, которые формируются в результате поисковых процессов в базе знаний концептов. Этот подход был первоначально предложен Рисбеком и Мартином [7].

Семантическая сеть или фрейм представляется в виде совокупности взаимосвязанных вершин и связей. Вершины представляют концепты, в то время как связи представляют отношения между концептами. Вершины иерархически организованы посредством связей типа IS-A, "намерение", "мотивация", "достижение"и др (см. рис.2). Установленные связи используются на этапе анализа в операции "передачи маркера"на семантической сети. Такой подход впервые был предложен в [2]. В последствии этот метод использовался в системах обработки естественно-языковых данных [4, 5].

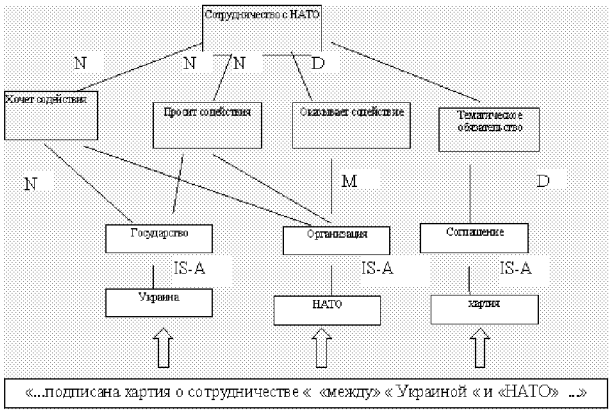

Рис. 2:

4 Представление языковых знаний

Все операции по поиску информации в памяти, актуализации концептов и выводу "умозаключений"осуществляется в системе КОНТЕНТ в пословном режиме.

Знания о словах в системе КОНТЕНТ находятся в словаре. Определение слова состоит из его значения и его синтаксической информации. Значение - это некоторый концепт или схема концептов определенных в модели предметной области.

Синтаксическая информация о слове включает такие понятия как: категорию, подкатегорию, тематические свойства, и собственно слово. Категория определяет, например, что "подписана"может использоваться как глагол (V). Подкатегория информирует нас о том, что "подписана"предшествует существительному (NP) "хартия".

5 Анализ

В результате пословного сканирования текста, посредством словаря, в котором определена связь слова с соот-

ветствующим концептом, инициируется агент концепта - одна из вершин вычислительной граф-схемы семантической сети. Агент концепта реализует две основных функции: передвигает маркер концепта по дуге графа; определяет состояние "активности"концепта. Управление основано на передвижении маркера по дуге вычислительной граф-схемы и определении состояния концептов в соответствии с установленными маркерами.

Операция вывода выполняется по каждому изменению состояния любой вершины схемы. Вывод строится на основании поиска активных вершин схемы. Например, установим маркер А на концепт "Соглашение", а маркер В на концепт "Сотрудничество". Далее, каждый из установленных маркеров должен передвинуться против связи типа "IS-A". После такого передвижения, необходимо проверить какие вершины содержат оба маркера. В приведенном примере - это концепт "Хартия". В данном случае мы можем утверждать, что "Хартия"является "Соглашением о сотрудничестве". Вывод образуется путем анализа атрибутов маркеров. С маркером связаны два основных атрибута: численное значение и адресный атрибут. Численное значение используется в представлении уровня значимости маркера. Адресный атрибут используется для идентификации маркера.

Конечный результат анализа - построение или отыскание в памяти понятийных структур или поиск в этих структурах с целью построения связей с другими структурами. Анализатор системы так же учитывает такие синтаксические категории как СУЩЕСТВИТЕЛЬНОЕ, ГЛАГОЛ, ИМЕННАЯ ГРУППА, ГЛАГОЛЬНАЯ ГРУППА, НАРЕЧИЕ и т.п., которые являются равноправными объектами концептуальной модели. Кроме этого применяется та же методика, при которой слова инициируют анализ концептов, а приписанные им агенты, находят и обследуют другие концептуальные структуры. Синтаксическая информация предложения используются в том случае, если только какой-либо семантический показатель, определенный в модели может быть определен путем использования синтаксической структуры. После этого синтаксическая информация "забывается"(см. рис. 3).

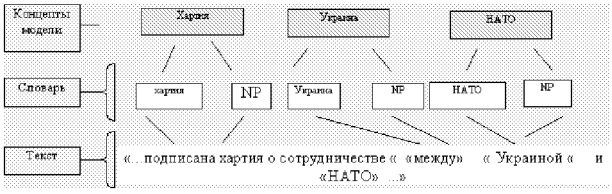

Рис. 3: Схема "чтения"текста

6 Представление "смысла "материала

"Смысл"прочитанного материала в системе КОНТЕНТ представляется так же в виде концептуальной схемы. В результате выполнения операции вывода, формируется семантическая схема, которая включают только активные вершины исходной концептуальной модели. Например в фразе, представленной выше в результате анализа состояния концептов модели были определены следующие концепты (см. рис. 4).

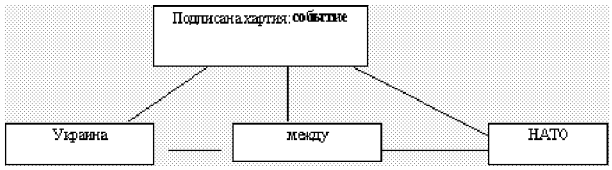

Рис. 4: Представление "Смысла"прочитанного материала

7 Обзор архитектуры системы КОНТЕНТ

Система КОНТЕНТ состоит из следующих основных взаимодействующих модулей:

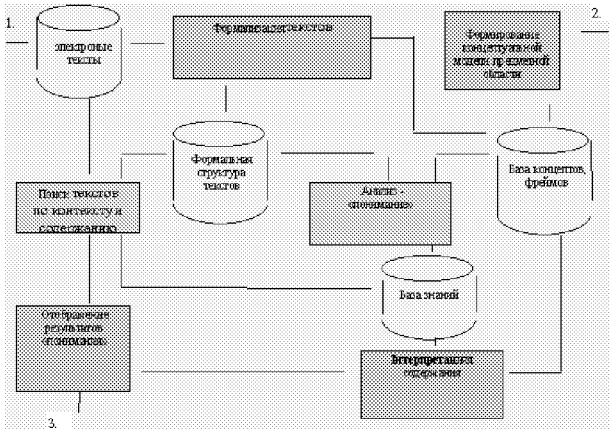

Рис. 5: Архитектура системы КОНТЕНТ

Модуль формализации теста, принимает на входе текст и порождает систему реляционных таблиц, содержащих формальную структуру текста. В дальнейшем все операции по анализу и "пониманию"текста осуществляются с использованием его формального представления средствами SQL.

Модуль представления знаний предметной области в виде концептуальной схемы. Основные функции модуля состоят в формировании модели понятийных знаний. Модель включает определения концепта или последовательности концептов и спецификацию связей между концептами.

Модуль анализа содержания. Осуществляет "чтение"текста и инициирование концептов в модели предметной области в соответствии с параметрами словоформ в предложениях.

Модуль интерпретации результатов анализа. Формирует семантические объекты в виде концептуальных зависимостей.

Модуль отображения и представления результатов анализа. Формирует графическое и табличное представление "смысла прочитанного"материала.

8 Заключение

Среди множества проблем, которые сопутствуют разработке следует выделить основные:

1) Надежное "понимание"текстового материала. Данная проблема может быть определена числом неправильных "догадок", которое система делает и степенью итогового понимания. Данная проблема связана с решением вопросов незнакомых слов, оговорок, нерелевантных междометий и др.

2) Скорость обработки текста обусловлена вопросами реактивности или производительности вычислительной системы. что обусловлено обработкой большого объема данных, хранимых в с относительно медленной памяти. Данная проблема решается путем распараллеливания обработки и доступа к данным.

Эти и множество других вопросов представляют объект текущих и дальнейших исследований в рамках данной работы.

Библиография

[1] Chang M. and Moldovan D. *Parallel Natural Language Processing on a Semantic Network Array Processor.*//"IEEE Transactions on Knowledge and Data Engineering", Vol.7,No 3, June 1995, pp. 391-405. .

[2] Charniak E. *A neat theory of marker passing*//Proc. Fourth Nat. Lng. Conf. Artificial Intelligence,1986. .

[3] Cowie J., Lehnert W. *Information Extraction* //Communications of the ACM.-1996.-Vol.39.-No.1.-P.80-91..

[4] Hirst G. *Semantic Interpretation and the Resolution of Ambiquity.* //Cambridge Univ.,1987. .

[5] Hu Y .H. and Simmons R. F. *Truly Parallel Understanding of Texp*//V joint Conf. Artificial Intelligence,1987. .

[6] Lehnert W. , Dyer M. , Johnson P. , Yang C. and Harley S. *BORIS - An Experiment in Depth Understanding of Narratives.* //. - "Artificial Intelligence", vol. 20, No 1, 1993. Pp. 15-62. .

[7] Riesbeck C.K., Martin C.E. *Direct memory acces parsing.* //Hillsdate, NJ: Lawrence Erlbaum,1977. –3.

[8] Schank R., Abelson R. *Scripts, Plans, Goals and Undestending*//Report 354. Yale Univ.1985. –3.

[9] Schank R., Birnbaum L. *Memory, meaning, and syntax.* //In "The Study of Language in the Cognitive Science". Cambridge, .–Ma: MIT Press, 1986. .

[10] Дерецкий В.А., Бородкина И.Л., Обуховская В.В., Богданова М.М., Ремарович С.С. *Один подход к построению компьютерных систем для анализа, обобщения и отображения содержания больших объемов текстовой информации.* //Сборник научных трудов ИПС НАНУ, Проблемы программирования, вып. 1, 1997 г. С.58-66. .