ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР СИСТЕМЫ ВОСТОК-О

О. Н. Очаковская

Вычислительный центр Сибирского отделения АН СССР
630 090 Новосибирск, СССР

Лингвистический процессор служит для ввода текстовой информации в экспериментальную информационную систему ВОСТОК-О, включающую логический вывод и встроенную простую модель времени (1), (2).

Процессор обладает следующими основными возможностями:

а) Производится анализ связного текста, состоящего из простых предложений,

б) Устанавливаются антецеденты простых анафорических ссылок (он, она, там),

в) Восстанавливается информация "по умолчанию".

Поступающая в систему информация преобразуется лингвистическим процессором в семантический граф, где все отношения являются предикатами первого порядка.

Процессор включает в себя базу данных, содержащую информацию о предметной области, и программную часть. База данных делится на две части:

- словарь, включающий морфологическую, синтаксическую и некоторую семантическую информацию о словах;

- библиотека, содержащая энциклопедическую информацию о понятиях, включенных в базу данных.

Процесс преобразования входного текста во внутреннее представление состоит из следующих подпроцессов:

(а) Извлечение конкретной информации, непосредственно содержащейся в тексте. При этом различается абсолютная и относительная информация,

(б) Пополнение объектов и отношений за счет информации "по умолчанию". Так, при анализе временных конструкций восстанавливаются отношения предшествования между последовательно упомянутыми событиями,

(в) Присоединение очередного сообщения к введенным ранее. При этом происходит отождествление вершин графа сообщения и основного графа. Этот процесс опирается на информацию о контексте.

Текст обрабатывается по предложениям. Каждое предложение порождает фрагмент семантического графа следующего вида: одну вершину - предикат и вершинны - объекты, соединенные с предикатом согласно модели управления данного предиката (хранится в базе данных) и подструктуру, состоящую из вершин-временных интервалов и связывающих их отношений. Причем глагол исходного предложения при анализе переходит в предикат, именные группы - в объекты и их характеристики, а обстоятельство времени - в подструктуру временных интервалов и отношений.

Вновь сформированный фрагмент присоединяется к уже созданному графу сообщения. При этом устанавливаются анафорические связи и склеиваются тождественные объектные вершины. Процесс объединения опирается на информацию о предшествующем тексте, которая накапливается в двух множествах (назовем их контекстными). В одном из этих множеств хранятся последние по порядку упоминания в тексте объекты, различающиеся семантическими типами и характеристиками. Они используются для идентификации ссылок на объекты по их типу и характеристике (этот студент). Второе множество используется для поиска антецедента анафорических ссылок, выраженных личными местоимениями (Она купила ее в магазине). В нем каждому объекту сопоставляется грамматический род породившей его именной группы, а также фиксируется порядок упоминания объектов в тексте. Отдельно хранится ссылка на элементы семантического графа, соответствующие времени действия последнего упомянутого события и его месту действия.

Лингвистический процессор особое внимание уделяет извлечению информации о времени. У обстоятельств времени допускается более сложное строение, чем у групп актантов (напр. "однажды зимой в течение 8 часов"). Кроме того, они подвергаются содержательной семантической интерпретации, использующей модель времени, в результате чего создается структура временных объектов и отношений (Т-структура).

Т-структура строится на основе информации как явно заданной в тексте (абсолютная информация): "вчера", "с 4 до 6", "3 дня назад, утром 20 марта", так и извлекаемой из сообщений посредством эмпирических правил, которые используют порядок анализируемых предложений и масштабность образуемых интервалов. Такая информация, получаемая "по умолчанию", считается относительной и вносится в семантическое представление лишь при отсутствии противоречащей ей абсолютной информации.

В заключение приведем пример фрагмента текста, анализируемого процессором:

"Позавчера днем профессор Терехов обедал с ассистентом Крыленко в кафе "Солнышко". Там профессор показал Крыленко незаконченную рукопись. Ассистент внимательно ее просмотрел, затем он отдал рукопись профессору".

Процессор реализован на языке сверхвысокого уровня СЕТЛ.

Обработка приведенного фрагмента требует около 10^7 опер. ЦП. Процессор использовался на 3-х русских текстах по 20÷30 предложений.

ЛИТЕРАТУРА

1. Е.Ю.Кандрашина, О.Н.Очаковская, Л.А.Голубева. Экспериментальная вопросно-ответная система, включающая простую модель времени с элементами логического вывода. В кн.: Взаимодействие с ЭВМ на естественном языке. Новосибирск: 1978, с. 207-222.

2. Kandrashina E.Yu., Ochakovskaya O.N. The Model of Understanding in the Project VOSTOK. In: Artificial Intelligence and Information-control Systems of Robots. Proc. of the Intern. Conf., Smolenice, CSSR, June 30 - July 4, 1980. p. 116-118.