

В.П. Гладун

Институт кибернетики им. В.М.Глушкова НАН Украины

glad@aduis.kiev.ua

1. Особенности метода.

Прежде всего, объясним содержащееся в заголовке статьи уточнение “в целенаправленных системах”. Процессы понимания естественно-языковых текстов в ситуациях целенаправленного поведения и в ситуациях, в которых отсутствует влияние цели, существенно различаются. В процессах целенаправленного поведения формирование семантической информации сопровождается фильтрацией ее по критериям адекватности имеющимся целям. Процессы понимания интегрируются с процессами решения задач. Это обстоятельство оказывает значительное влияние на методы синтаксического и семантического анализа.

Естественно-языковое общение с компьютером при формировании и решении задач всегда считалось престижной целью исследований в области искусственного интеллекта. Во многих практических ситуациях удается уменьшить сложность проблемы путем использования жестких синтаксических и семантических ограничений на допустимые естественно-языковые тексты. Проект, о котором идет речь в этой статье, прежде всего, ориентирован на ситуации, в которых есть возможность упростить семантический анализ за счет использования сведений о тематической направленности текстов, но нет возможности ограничить естественно-языковое общение фиксированным набором естественно-языковых фраз. Такой характер имеют, например, задачи диспетчеризации ремонтных работ по естественно-языковым заявкам.

При анализе естественно-языковых текстов базовой операцией является распознавание синтаксических и семантических отношений, связывающих слова текста. Распознавание отношений осуществляется на основе их описаний (моделей). Такого рода модели обязательно присутствуют во всех методах анализа, хотя не всегда явно. В большинстве методов анализа процессу распознавания отношений предшествует перевод исходного естественно-языкового представления распознаваемых объектов (отношений) в язык категорий традиционной грамматики (число, род, падеж, время и т.д.). Правила распознавания синтаксических и семантических отношений оперируют грамматическими описаниями слов. Привязка к грамматическим описаниям элементов текста влечет следующие недостатки: разнородность процессов обработки отдельных слов и словосочетаний; громоздкость процессов обработки; сложность адаптации к изменениям лексики и предметной области пользователя; трудоемкость разработки. Между тем, переход

к грамматическим описаниям не является обязательным условием для выполнения анализа естественно-языковых текстов. Информация, необходимая для распознавания синтаксических и семантических отношений, содержится непосредственно в тексте. Доказательством тому служат "человеческие" процессы анализа естественно-языковых текстов, не связанные с грамматическими категориями и правилами. Поэтому правомочен другой подход, основанный на использовании соответствий между отношениями и средствами их выражения в естественно-языковых текстах. В силу своих принципиальных особенностей такой подход позволяет исключить названные выше недостатки.

Модели отношений, в которых для распознавания синтаксических и семантических отношений используются элементы естественно-языковых текстов, в основном морфы, назовем моделями непосредственного распознавания

Опишем принципы анализа текстов на основе моделей непосредственного распознавания [1].

2. Знания о языке.

Синтаксическое представление текста (СИНП) строится из структур, представляющих синтаксические связи между знаменательными словами текста. Во флективных языках синтаксические отношения знаменательных слов кодируются сегментами словосочетаний, образующимися в результате исключения из словосочетаний основ знаменательных слов. Назовем эти части словосочетаний синтаксическими определителями. Например, в словосочетаниях "толстая книга", "машина движется", "двигатель не работает" синтаксическими определителями являются: -ая -а; -а -ется; -* (нуль-флексия) не -ет.

Синтаксические определители можно использовать для распознавания синтаксических связей между знаменательными словами. Однако однозначность распознавания иногда не достигается из-за омонимии определителей - существования пересечений множеств определителей, кодирующих разные синтаксические отношения. Например, из-за омонимии определителей оказываются неразличимыми синтаксические отношения, соответствующие словосочетаниям "догоняла собака" и "догоняла человека". В таких случаях необходимо иметь дополнительную информацию о словосочетании, обеспечивающую однозначное распознавание синтаксического отношения. В связи с этим основы знаменательных слов группируются в семантические классы, и каждому синтаксическому отношению сопоставляется описание вида $\{ \langle D_1^a, k\phi_1, k\phi\phi_1 \rangle, \langle D_2^a, k\phi_2, k\phi\phi_2 \rangle \dots \langle D_i^a, k\phi_i, k\phi\phi_i \rangle \dots \}$, где a -имя отношения, D_i^a - множество определителей, кодирующих это отношение в случае, когда основы, являющиеся его аргументами, относятся к классам $k\phi_i$ и $k\phi\phi_i$. Кортежи $\langle D_i^a, k\phi_i, k\phi\phi_i \rangle$ называются синтаксическими корреляторами. Однозначное распознавание синтаксического отношения достигается, если для любых двух синтаксических корреляторов $\langle D^a, k\phi, k\phi\phi \rangle$ и $\langle D^b, k\phi, k\phi\phi \rangle$, описывающих разные отношения a и b , но включающих общую пару классов основ $k\phi$ и $k\phi\phi$ выполняется условие $D^a \cap D^b = \emptyset$.

Классы основ определяются семантическими признаками (действие, состояние, качество, вещество, среда, причина и т.п.) или сочетаниями семантических признаков. Для распознавания синтаксического отношения нужно построить описание анализируемого словосочетания в виде $\langle d, k\phi, k\phi\phi \rangle$, где d - выделенный из словосочетания синтаксический определитель, $k\phi$ и $k\phi\phi$ - классы основ знаменательных слов, входящих в словосочетание. Классы основ определяются с помощью словаря основ, содержащего основы с указанием их семантических признаков. Построенное описание словосочетания сопоставляется с синтаксическими корреляторами, входящими в описания синтаксических отношений. В данном случае термин "синтаксические отношения" весьма условен, так как отношения, выделенные таким образом, не только определяют в совокупности синтаксическую структуру текста, но и раскрывают его поверхностную семантику.

3.Алгоритм анализа предложения.

Алгоритм анализа задает порядок анализа предложений с целью выделения словосочетаний, которым соответствуют синтаксические отношения. В результате строится СИНП, отображающее дерево подчинения предложения. В основе принятого порядка анализа словосочетаний лежат закономерности структуры предложений русского языка.

Для отбора фраз, имеющих отношение к задаче, используются так называемые *целевые отношения*, которые определяют целевую ситуацию. Например, для выяснения позиций депутатов в решении некоторой государственной проблемы в их программах отбираются фразы, соответствующие целевому отношению "X выступает заY" (X-идентификатор депутата, Y-вариант решения). В естественно-языковых фразах это отношение представляется словами "выступает за", "поддерживает", "голосует за" и т. п. Именно с таких слов и словосочетаний, коль скоро их удастся отыскать в естественно-языковой фразе, начинается ее разбор. В дальнейшем анализу подвергаются части предложения, непосредственно примыкающие к найденному словосочетанию, в связи с чем оно называется *опорным*.

Расширение обработанной части предложения осуществляется путем присоединения примыкающих к ней знаменательных слов. Знаменательное слово, примыкающее к обработанной группе слов, присоединяется к группе, если в группе имеется знаменательное слово, связанное с ним каким-либо синтаксическим отношением. Для распознавания отношения, которое связывает анализируемую пару знаменательных слов, выполняется их морфологический анализ, в результате которого выделяются основы и синтаксический определитель. Разделение словоформ на основы и флексии производится алгоритмами обращения к словарю, в котором хранятся основы и синтаксические определители. Описание анализируемой пары сопоставляется с синтаксическими корреляторами. Как только отыскивается словосочетание, для которого удастся распознать синтаксическое отношение, строится СИНП, представляющее это словосочетание.

Если в группе нет знаменательных слов, связанных с примыкающим к группе словом, в остальной части предложения отыскивается новое опорное словосочетание, в качестве которого выбирается ближайшее синтаксически связанное словосочетание двух соседних знаменательных слов. Начиная с этого словосочетания, строится СИНП новой группы слов. Распознавание синтаксических отношений осуществляется стандартным для данного алгоритма способом - путем сопоставления описаний словосочетаний с синтаксическими корреляторами.

После выделения новой группы слов предпринимается попытка найти в ней знаменательное слово, образующее синтаксически связанное словосочетание с каким-либо знаменательным словом разобранной группы слов, расположенной в предложении непосредственно слева. Если поиск словосочетания, связывающего группы, оказывается успешным, синтаксические представления групп связываются синтаксическим отношением, соответствующим найденному словосочетанию. Объединенная группа слов максимально расширяется влево и вправо. Процесс формирования и слияния групп продолжается, пока не окажется разобраным все предложение.

Предпринимаются меры, уменьшающие переборы при поиске связывающих словосочетаний во время расширения или объединения групп, а также препятствующие выделению ложных связывающих словосочетаний.

В качестве примера рассмотрим процесс анализа фразы "При устранении внешнего поля подсистема расщепленных уровней исчезает". Прежде всего, осуществляется просмотр фразы слева направо с целью выделения первого опорного словосочетания. Допустим,

таким словосочетанием является словосочетание "внешнего поля". Далее предпринимается попытка расширить найденную группу слов. В результате к опорному словосочетанию присоединяется слово "устранении", образующее синтаксически связанное словосочетание со словом "поля". Новую разобранную группу слов "устранении внешнего поля" расширить не удастся, так как единственное примыкающее к ней знаменательное слово "подсистема" синтаксически связанных словосочетаний со знаменательными словами группы не образует. Поиск нового опорного словосочетания приводит к выделению словосочетания "расщепленных уровней", на основе которого в дальнейшем формируется группа "подсистема расщепленных уровней исчезает". После этого во фразе не остается знаменательных слов, не входящих в какую-либо разобранную группу. Теперь нужно соединить синтаксические представления образовавшихся групп. Это осуществляется с помощью словосочетания "при устранении исчезает", включающего по одному знаменательному слову из каждой группы.

С учетом ограничений, препятствующих выделению ложных словосочетаний, при поиске словосочетаний не рассматриваются пары слов ("внешнего", "подсистема"), ("расщепленных", "исчезает"), ("устранение", "расщепленных"), ("устранении", "уровней"), ("внешнего", "расщепленных"), ("внешнего", "уровней"), ("внешнего", "исчезает"), ("поля", "расщепленных"), ("поля", "уровней").

4. Использование естественно-языкового лингвистического процессора в системах принятия решений.

Критерием понимания естественно-языкового сообщения системой, решающей задачи, является адекватность ее ответной реакции [2]. В связи с этим возникает проблема установления соответствия между естественно-языковыми фразами и реакциями системы.

В настоящее время создан программный естественно-языковый лингвистический процессор, работающий на основе описанных выше методов. Ведутся работы по интеграции его с системами принятия решений.

СИНП, формируемое при анализе предложения, ограничивается по составу и числу отношений в "левой" (слева от целевого отношения) и "правой" (справа от целевого отношения) его частях. В результате введенных ограничений СИНП предложения приобретает одну из стандартных форм, которые используются для выбора действий при решении задач.

Например, в естественно-языковых системах диспетчеризации ремонтных работ по построенным синтаксическим представлениям фраз выбираются решения по комплектованию ремонтных бригад. В системе планирования политических акций [3] стандартные типы СИНП входят в состав продукционных правил, используемых для принятия политических решений.

Приведенные начальные эксперименты свидетельствуют о работоспособности предложенных методов, но пока еще не позволяют в достаточной мере оценить их эффективность. Развитие подхода связано, прежде всего, с расширением и совершенствованием лингвистической базы знаний.

Литература

Гладун В. П. Процессы формирования новых знаний.-София: Педагог, 1994.-190с.

Файн В. С. Машинное понимание естественного языка в рамках концепции реагирования// Интеллектуальные процессы и их моделирование.-М. Наука, 1987.-С.375-392.

Gladun V., Vaschenko N., Reutov G. Computer systems supporting political planning// IFIP Congress '94, Hamburg, Germany, 1994.-Elsevier Science B. V. (North-Holland), 1994.-P.39-44.