УДК 519.7663:768
С.Ж.КАРАБАЕВА, С.С.РЫСПАЕВА, Ж.С.МАМЫТБЕКОВА
E-mail: ksucta@elcat.kg.

КОМПЬЮТЕРНЫЙ АНАЛИЗ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Бул макалада тексттерди компьютердик иштеп чыгууда математикалык лингвистикалык ыкмалар каралган.

В данной статье рассмотрены методы математической лингвистики, применяемые в компьютерной обработке текстов.

This article is devoted to the mathematical methods of linguistics applying in computer processing of texts.

В этой статье описывается исследование в области Искусственного Интеллекта по порождению естественных языков, при этом особое внимание уделяется конкретным проблемам, которые требуют разрешения.

В настоящее время расширяются возможности использования компьютерной технологии в различных сферах деятельности, в том числе в образовании. Достижения в области информатики и вычислительной техники вызвали изменения в содержании и методах преподавания многих вузовских дисциплин. Бурное развитие информационных технологий и их внедрение во все сферы жизни вносят изменения в содержание различных видов деятельности. Во всех развитых и во многих развивающихся странах идут интенсивные процессы информатизации образования. Разрабатываются пути повышения результативности общего образования, вкладываются большие средства в разработку и внедрение новых информационных технологий.

Доступные сегодня вычислительные мощности позволяют применять для обработки больших массивов документов широкий класс математических методов, способствующих эффективному решению задач поиска, классификации, кластерного анализа, выявления скрытых закономерностей в данных и др.

Формальные языки, так же как и языки естественные, можно рассматривать с точки зрения их формы, структуры, иначе говоря, синтаксиса, и с точки зрения смысла, вкладываемого в предложения языка, то есть семантики. Наиболее простым и поэтому наиболее проработанным является синтаксическое описание языка, то есть его грамматика.

Порождение текстов на естественном языке – процесс преднамеренного построения текста на естественном языке с целью решения определенных коммуникативных задач. Термин «текст» рассматривается как общий, рекурсивный термин, который может относиться к письменному или устному высказыванию, или к отдельным частям высказывания. При порождении текстов, в устной или письменной форме, человеку важно обдумать и отредактировать производимое

высказывание. Едва ли можно сказать, что большинство программ может «говорить» сегодня, в основном все они лишь выводят слова на экран. Так как для программы порождения текстов на сегодняшний день не стоит вопрос конструирования фразы, эти детали принимаются во внимание только тогда, когда они задействованы в создании программы.

В отличие от организации процесса понимания, который, на первый взгляд, может следовать традиционным стадиям лингвистического анализа: морфология, синтаксис, семантика, прагматика, процесс порождения имеет существенно отличный характер. Этот факт следует непосредственно из присущих различий в информационном потоке в двух процессах.

Порождение имеет противоположный информационный поток. Оно переходит от содержания к форме, от целей и перспектив к линейно упорядоченным словам и синтаксическим маркерам. Большинство систем порождения производит поверхностные тексты последовательно слева направо, но только приняв решение сверху-вниз по содержанию и форме текста в целом. Проблема генератора состоит в том, чтобы выбрать из поставленных источников, как правильно сообщить о желаемых умозаключениях аудитории и какую информацию опустить из явного упоминания в тексте.

Можно вообразить, что процесс порождения так же организован, как и процесс понимания, только в противоположном порядке. К некотором смысле это верно: идентификация намерения (цели) в значительной степени предшествует любой детализации информации, которая предназначается для аудитории: планирование риторической структуры, например, в значительной степени, предшествует любой синтаксической структуре, а синтаксический контекст слова должен быть зафиксирован, прежде чем будут известны морфологическая и суперсегментная формы, которые примет слово.

Синтаксис и словарь языка становятся как ресурсами, так и ограничениями, определяя элементы, доступные для создания текста, а также зависимости между ними, которые определяют возможные правильные комбинации. Эти зависимости и тот факт, что они по умолчанию управляют, когда информация, от которой зависит каждое решение, становится доступной, – основная причина, почему программы порождения в значительной степени следуют стандартным стадиям, определенным лингвистами. Компоненты порождения естественного языка не существуют сами по себе. Они расположены внутри человеко-машинного интерфейса, который также используют и компоненты понимания естественного языка, – ВВОД в систему. В хорошем человеко-машинном интерфейсе сегодня также хотелось бы видеть координированную графическую поддержку ввода и вывода, дополняя систему ВВОДа-ВЫВОДа естественного языка. Интерфейс может закончиться здесь, а может также включать в себя другие общедоступные компоненты, типа контроллера дискурса, который указывает генератору, какие действия нужно предпринять, а также координирует интерпретации, сделанные компонентом понимания.

Сегодня большинство исследователей в этой области работает, в основном, с экспертными системами, где процесс общения контролируется программой, а не пользователем. Кроме того, ЭС

и интеллектуальные машинные обучающие программы, вероятно, способны понимать довольно сложные тексты, что делает их привлекательными для специалистов, готовых работать с уже разработанными системами.

Цели должны обычно передавать некоторую информацию аудитории или побуждать ее к действиям или рассуждениям. Социальные и психологические, а также практические мотивы, побуждающие человека к общению, естественно, неприменимы для сегодняшних компьютерных программ. Планирование включает в себя отбор (преднамеренное вычеркивание) информационных модулей, которые появляются в тексте (например, концепции, отношения, индивидуальность).

При автоматической обработке текста возникает проблема “новых“ слов. Для синтаксического анализа и синтеза необходимо знать грамматические характеристики слов. Если слова в словаре нет, то морфологический анализ не может быть выполнен, а следовательно, не могут быть определены грамматические характеристики слова.

Если по текстам большого объема составить словарь словоформ и назначить каждой словоформе некоторые грамматические признаки, а затем преобразовать данный словарь в обратный словарь словоформ, то можно обнаружить, что многие участки словаря имеют одинаковые наборы признаков.

Обратный словарь используется для автоматического морфологического анализа текстов, если составляющие их словоформы отождествлять со словоформами словаря и приписывать им грамматическую информацию, указанную в словаре. Словоформам текста, которые не находятся в словаре, можно приписывать грамматическую информацию тех словоформ словаря, концы которых в максимальной степени совпадают с концами этих новых словоформ текста.

Объем обратного словаря можно сократить, если на всех его участках оставить по две словоформы: начальную и конечную. Более того, из этих двух словоформ можно оставить только одну, и если словоформа текста не совпадет ни с одной словоформой обратного словаря, то ей приписывается информация непосредственно предшествующей словоформы этого словаря.

Синтаксический анализ формальных языков во многом напоминает известный по школе грамматический разбор предложений.

Анализ текста — процесс получения высококачественной информации из текста на естественном языке. Как правило, для этого применяется статистическое обучение на основе шаблонов: входной текст разделяется с помощью шаблонов, затем производится обработка полученных данных.

Факт и явления языка и речевой деятельности имеют разные признаки и поэтому могут быть рассмотрены с разных сторон. Так, содержание предложения Всякий равносторонний треугольник есть равноугольный треугольник логик определит как суждение тождества, состоящее из субъекта (S = всякий равносторонний треугольник), связки и предиката (S = равноугольный треугольник), причем будет отмечено, что субъект и предикат имеют один и тот же объем понятия, поскольку обозначают понятия равнозначащие. Рассмотрим пример.

$G_1 = (Z_1, P_1, V_1, N_1)$.

$P_1 : Z_1 --> \{0 \mid 1\}$ *или*

$Z_1 --> 0 \mid 1\, Z_1 0 \mid Z_1 1$

Эта грамматика регулярна и порождает множество целых двоичных чисел. Грамматика $G_2 = (Z_2, P_2, V_2, N_2);\ P_2 :\ \ Z_2 --> Z_2 W \mid W;\ \ W --> 0 \mid 1$ по виду правил является контекстно-свободной, но порождает она то же множество чисел, что и G_1.

При разборе предложения обнаруживаются два аспекта анализа – логический и грамматический.

Синтаксический анализ является очень важным приложением Пролога и логического программирования. В действительности, происхождение Пролога связано с попыткой использовать логику для выражения грамматических правил и формализации процесса синтаксического разбора. Во многих грамматиках используются рекурсивные правила.

Логические грамматики превратились с течением лет в инструментарий высокого уровня, и теперь они позволяют пользователю сконцентрироваться на лингвистических феноменах. Грамматики, построенные на определенных предложениях, поддерживают использование логики для обработки данных естественного языка, и они подготовили почву для практической работы лингвистов на языке программирования PROLOG.

Понятие грамматик, построенных на определенных предложениях (DCGs), как особого случая метаморфозных грамматик было введено в 1978 году Перейрой и Уорреном в качестве грамматического формализма, для которого PROLOG имеет эффективный механизм синтаксического анализа. Одни практические системы были созданы для одновременного использования синтаксического и семантического знания для привнесения логики в структуру, содержа в себе информацию для семантической интерпретации. Другие системы были выстроены на более чем одном уровне трансляции; использование синтаксического и семантического знания осуществлялось отдельно друг от друга, и конечным результатом являлось в PROLOGe предложение Хорна, выполнение которого осуществлялось механизмом планирования (qv).

Грамматики описывают структуру (синтаксис) языков множеством продукций (правил, перерабатывающих текст). Например, правилом

sentence --> noun-phrase verb-phrase

устанавливается связь между тремя нетерминальными символами: предложение может состоять из именной группы и следующей за ней глагольной группы.

Такие правила могут быть отображены в PROLOGe следующим образом:

```
sentence (S1, S3): -
noun-phrase (S1, S2),
verb-phrase (S2, S3).
    verb-phrase (S1, S2): -
    connects (S1, writes, S2).
    connects (1, each, 2).
```

connects (2, author, 3).

connects (3, writes, 4).

(Примечание: предикаты (т.е. выражения с неопределенными терминами или переменными, которые преобразуются в истинные или ложные высказывания при выборе конкретных значений для этих самых терминов, заносятся в PROLOG через запятую. Переменные отличаются от констант первой заглавной буквой.)

Чтобы проверить правильность построения предложения, необходимо указать цель

? - sentence (X, Y

(где ? - бинарное обозначение структуры (или бинарный функтор), содержащееся в любой системе PROLOG) и продемонстрировать, что она подтверждается предыдущими условиями. Используя список в качестве информационной структуры для представления предложения, числа больше не нужны, так как PROLOG имеет устройство синтаксического анализа, способного перевести:

? - sentence ([each, author, writes]. [ ]).

Грамматики, построенные на определенных предложениях, являются объемом понятия контекстно-свободных грамматик, которые также могут быть транслированы на язык PROLOG. Грамматики, построенные на определенных предложениях, позволяют любому логическому выражению стать нетерминальным, они построены на логических символах: константах, переменных, выражениях, а не только на одних константах. Также они имеют только один нетерминальный символ в левой части каждого правила. Контекстные зависимости (контекстные отношения подчинения) описываются логическими переменными в рамках параметров (или независимых переменных) грамматических символов.

Каждое грамматическое правило, типа: p(X) --> q(X), получает группу входящих данных, анализирует некую исходную часть и генерирует остаток для дальнейшего анализа.

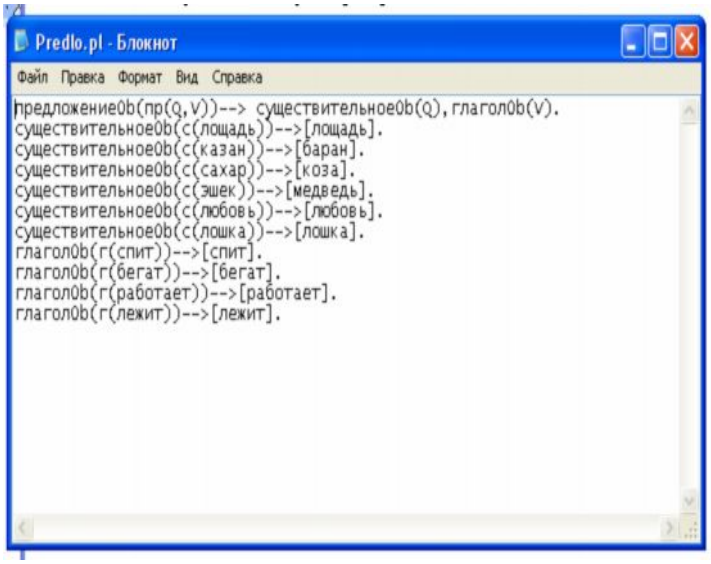

Рис.1. Морфологический анализ в блокноте

Рис.2. Морфологический анализ на Прологе

Структура иерархии грамматических понятий, являющаяся результатом разбора, в теории формальных языков носит название дерева синтаксического разбора. Листьями этого дерева являются конкретные слова языка или терминалы.

В грамматике, представленной программой, сначала обрабатываются группа существительных с произвольно большим количеством прилагательных, а затем рассматривается группа глагола.

Задачи грамматического разбора предложений языка или, иначе, синтаксического анализа, связаны с построением дерева вывода, определением последовательности применения правил, обеспечивающей соответствие входного предложения и начального символа грамматики.

СПИСОК ЛИТЕРАТУРЫ

1. Стерлинг Л., Шапиро Э. Искусство программирования на языке Пролог. – М.: Мир, 1990.

2. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. – М., 1985.

3. Заболеева_ А.В., Зотова, Камаев В.А. Лингвистическое обеспечение автоматизированных систем. – М.: Высшая школа, 2008.