Синтез текстовой информации на английском языке при решении задач дистанционного обучения иностранному языку с использованием объектного подхода.

Н.В. Крапухина[1], С.Ю. Кулехин[2]

В работе представлено описание подсистемы синтеза естественно-языковых текстовой информации, ориентированной на диалог и реферирование в основе которой, лежит представление текста в виде объектно-ориентированной семантической сети, генерируемой с использованием фреймов предложений на базе опорных точек.

The description of system oriented to the dialog and reviewing. This system adapts the natural linguistic texts and is based on the presentation of the text as the object-oriented semantic net, which is generated using the frames of the sentences based on the supporting points.

Введение.

Многообразие функций, выполняемых современными прикладными лингвистическими моделями, нашло отражение в понятии автоматизированной системе обработки текстов (АСОТ). «Такие системы моделируют способность человека понимать текст (это анализ), а также хранить и перерабатывать информацию на некоторых внутренних промежуточных семантических метаязыках (языках представления знаний)» [Городецкий 1983, 9]. Близкое представление вкладывается также в понятие лингвистического автомата. «Под лингвистическим автоматом понимается комбинация ЭВМ и работающих алгоритмов и программ, предназначенных для автоматической переработки текстов» [Лингвистические проблемы 1980, 171]

Вся история компьютерного моделирования естественного языка, насчитывающая уже несколько десятилетий, приводит к выводу о том, что свойства языка вызывают отторжение формализации на ЭВМ. Попытки создать машинные аналоги естественного языка, связанные с появлением компьютеров, потребовали разработки теоретической базы для адекватного моделирования человеческого языка, что вызвало интенсивное развитие узкоспециализированных работ как теоретического, так и прикладного характера (формальные грамматики и автоматизированные словари различного типа для конкретных языков и т.п.).

В конце 60-х годов в системах ИИ сформировалось направление разработки ЕЯ-систем. Эти системы предназначены для разработки систем, реализующих процесс общения на естественном языке.

Ограничение предметной области имеет принципиальное значение, т.к. позволяет наложить ограничения на лексику, семантику и синтез языка. При этом ЕЯ система должна допускать возможность изменения своих знаний, зависящих от специфики рассматриваемой предметной области.

[1] 117936, Москва, Ленинский пр-т д. 4 МИСиС(ТУ) Krapuhina@misis.ru
[2] 117936, Москва, Ленинский пр-т д. 4 МИСиС(ТУ) Serg@planetashop.ru

В описываемой работе предметная область имеет локальное значение. Т.е. система автоматически анализирует «предложенный» ей текст и строит базу знаний на основе заложенной в данный текст информации.

Описываемая в данной работе подсистема является составляющей новой АСОТ «iSemNet» представленной в виде 4-х компонентов:

- компонент анализа текстовой информации на английском языке
- компонент синтеза текстовой информации на английском языке (описывается в данной статье)
- компонент сценариев
- компонент хранения и обработки семантической информации.

АСОТ «iSemNet» создается на кафедре Инженерной Кибернетики МИСиС(ГУ) группой аспирантов под руководством заведующей кафедры Крапухиной Н.В.

Функциональная структура системы.

При разработке функциональной схемы iSemNet были выделены следующие составляющие процесса анализа и синтеза текстов: анализ входящего текста, генерация семантической сети, синтез предложений по сети. В системе использован комплексный подход к построению структуры базы знаний. Динамическая подсистема базы знаний iSemNet относится не к какой-либо отдельной модели, а представляет собой интеграцию целого ряда способов представления знаний и методов искусственного интеллекта по обработке знаний:

- Семантические сети. Ими представлены знания о тексте. Новизной данной работы является то, что узлами сети (направленного графа) являются объекты текста (объекты, которые производят действие и субъекты, над которыми совершаются действия), причем субъект может быть скрытым. В качестве узла может выступать также любая составляющая сети (сказуемое, свойство и т.п.). Данное свойство обусловлено семантической сложностью естественно-языковых текстов, а реализация стала возможной благодаря объектно-ориентрованному подходу представления текстов. Ребрами графа являются глаголы (сказуемые) предложений. Все объекты сети (узлы и ребра) имеют свои свойства (прилагательные, числительные, наречия и др. лексикографические единицы).
- Фреймы. Фреймами представлены структуры предложений английского языка. По этим структурам была составлена БЗ фреймов. Фреймы обладают свойством неограниченной вложенности, что позволяет описывать ими практически любое английское предложение. Данный метод давно и небезуспешно применяется для хранения структур английских предложений, однако в данной работе он был дополнен свойством неограниченной вложенности и динамического расширения базы данных фреймов в системе обучения системы.
- Продукционные правила. По этим правилам трансформируются слова в нужные формы, проводится лексико-грамматический анализ, а также правила используются при составлении предложений. Продукционные правила имеют вид:

  $W_n = F\ (W, \{S_i\}, x_1, x_2, \ldots, x_n)$,

где

W_n – требуемое слово (словосочетание) в нужной форме,

F – процедура поиска/преобразования слова(словосочетания),

W – слово в простой форме (инфинитив глагола в настоящем времени, существительное в единственном числе именительном падеже и т.п.)

$\{S_i\}$ – семантическая сеть по данному текущему тексту,

$x_1, \ldots, x_n$ – дополнительные параметры, позволяющие однозначно определить форму трансформируемого слова (-сочетания).

- Мультипарсинг - метод обхода семантической сети. Данный метод оригинален тем, что позволяет параллельно «вести» по графу несколько маркеров и, тем самым, распараллеливать процесс поиска информации в семантической сети, что увеличивает скорость поиска.

Выбранные методы позволяют эффективно реализовать процесс синтеза текстовой информации. В настоящее время семантические сети стали одной из самых распространенных тем в теории искусственного интеллекта, применительно к лингвистике. Это связано с рядом причин:

- гибкостью представления знаний;
- разнообразием видов семантических сетей;
- легкостью манипулирования;
- наглядностью изображения.

По этим причинам в разрабатываемой системе основная часть интеллектуальной подсистемы а именно, динамическая часть БЗ – выполнена в виде семантической сети.

Объектный подход и многоуровневый анализ текстовой информации.

Особенностью данной работы является, использование создаваемого с 60-х годов объектно-ориентированного подхода в программировании в области компьютерной лингвистики. Проведенный синтаксический анализ английских текстов и объектно-ориентированных технологий представления моделей, позволил авторам разработать и применить в алгоритмах и программе свой подход к формальному представлению знаний о текстах.

Создаваемая модель текста или объекта содержит не все признаки и свойства представляемого ею предмета (понятия), а только те, которые присутствуют в рассматриваемом тексте. Тем самым модель "беднее", а, следовательно, проще представляемого ею предмета (понятия). Но главное в том, что модель есть формальная конструкция: формальный характер моделей позволяет определить формальные зависимости между ними и формальные операции над ними. Это упрощает как разработку и изучение (анализ) моделей (текстов), так и их реализацию на компьютере. В частности, формальный характер моделей позволяет получить формальную модель рассматриваемого текста как композицию формальных моделей ее компонентов (субъектов, объектов, их отношений и свойств и др.).

Семантическая сеть реализует полиморфизм (одно из главных свойств ООП) в двух плоскостях (полиморфный полиморфизм):

1. «горизонтальный» полиморфизм – в сети существуют объекты трех типов – «сущности», «связи», «свойства». В то же время все объекты сети равнозначны, например при поиске по сети. Также при необходимости возможна реализация мутации объектов одного типа в другой.
2. «вертикальный» полиморфизм – каждый объект сети может содержать в себе подсеть, функционально равнозначную сети верхнего уровня. Принципиальных ограничений ни на глубину вложенности, ни на размеры вложенных сетей нет.

Полиморфность объектов сети позволяет с помощь нее описывать реальные тексты произвольного объема и сложности с неограниченной детализацией объектов.

Ядром системы анализа является преставление модели анализируемого текста в виде семантической сети. Этим самым устраняется всякое влияние конкретного естественного языка, его отличительные особенности.

Но перед тем как достигнуть уровня семантики текста необходимо осуществить его анализ на предыдущих уровнях – морфологическом, синтаксическом.

Поэтому в модуле анализа обработка текстовой информации осуществляется поэтапно:

- синтаксический анализ;
- семантический анализ.

Цель синтаксического анализа – автоматическое построение функционального дерева фразы, т.е. определение взаимозависимостей между разно уровневыми элементами предложения. В данном случае, синтаксический анализ идет с опорой на служебные слова и набор правил построения английского предложения. Это позволяет при минимальном наборе служебных слов (порядка ста) анализировать тексты с буквально неограниченной лексикой, ибо неизвестное слово в противоположность многим другим методам не является препятствием данному алгоритму.

Реализованный алгоритм в основе своей опирается на идеи представления предложения в виде деревьев синтаксического подчинения. Отличие состоит в методе поиска ключевых элементов предложения. При их поиске опора делалась, в первую очередь, на формальные, внешние признаки, (т.н. «опорные точки»), а потом уже на место в предложении.

Этот метод анализа текстовой информации впервые был предложен Л.В.Щербой, З.М.Цветковой, В.И.Ноткиной и развит В.В.Милашевичем, Е.П. Грединой.

Основной задачей синтаксического анализа является отыскание в предложении его главных членов - сказуемого, субъекта и объекта. Собственно анализ начинается с поиска сказуемого, как центрального звена английского предложения, найдя которое можно определить структуру всего предложения в целом. Далее анализируются найденные комплексы субъекта и объекта – определяется их состав и структура связей между элементами. При этом учитываются как одноранговые отношения, так и отношения принадлежности одного элемента другому.

Следующий этап анализа – семантический, результаты работы которого представляются в виде семантической сети. Она является наиболее простым и универсальным средством представления знаний в системах искусственного интеллекта и представляет собой ориентированный граф, вершины которого обозначают сущности (объекты), а ребра – отношения (связи) между ними. Имена вершин и ребер совпадают с именами соответствующих сущностей и отношений, используемыми в естественном языке. Ребро и две связываемые им вершины представляют минимальную смысловую информацию – факт наличия связи определенного типа между соответствующими объектами.

Описание модуля синтеза осмысленной текстовой информации.

Модуль синтеза текстовой информации на английском языке осуществляет обратную связь с пользователем АСОТ. Отметим, что термин «синтез текста» употребляется обычно в отношении письменного текста, для устной формы функционирования языка обычно употребляется термин «синтез (звучащей) речи». В зависимости от реализуемой задачи, модуль синтеза может использоваться для разных целей:

- в задачах автоматического реферирования - компонент генерации предложений рефератов;
- в задачах обучения – компонент ведения диалога с пользователем на тему какого-то заданного текста (генерация вопросов и ответов);

Процесс генерации состоит из 2-х составляющих:

1. Генерация смысла высказывания
2. Синтез высказывания на естественном языке

Результатом выполнения 1-го этапа является внутреннее представление смысла генерируемого высказывания. При этом решаются следующие подзадачи:

- выделяются аспекты, интересующие пользователя
- определяется информация, которая должна быть сообщена пользователю.

Поиск информации осуществляется обходом построенной семантической сети в соответствии с заданным вопросом (мультипарсинг). Анализатор находит ту составляющую, о которой спрашивает пользователь и передает ссылку на эту составляющую в блок синтеза.

На втором этапе решаются следующие подзадачи:

- построение синтаксиса структуры отдельных предложений
- морфологический синтез словоформ.

В соответствии с заданным вопросом и найденным ответом блок синтеза выбирает из базы знаний фрейм ответа. Затем, проходом по семантической сети, заполняются необходимые поля фрейма. Фреймы предложений являются динамическими и адаптируемыми под текущую «беседу».

Многие системы автоматического синтеза текстовой информации опираются на некий словарь, изначально присутствующий в системе. И способность таких систем генерировать новую информацию ограничивается полнотой и размерностью такого словаря. Хорошо, когда такие системы являются самообучаемыми или пополняемыми – тогда есть возможность постоянно расширять словарь и – как следствие - область применения системы. Но такой подход вызывает необходимость постоянно пополнять словарь, искать новые физические хранилища постоянно увеличивающейся словарной базы, что отрицательно сказывается

на возможности переноса системы и ее распространении. Одной из особенностью данной работы является то, что словарь состоит из некоторого ограниченного набора служебных слов, основных глаголов и других синтаксических элементов, которые не требуется постоянно пополнять. Опираясь на такой словарь и любой текст, на основе которого ведется обучение в текущем сеансе, модуль может генерировать необходимые в текущий момент обучения предложения.

Особенность данной работы и ее практическая ценность

Особенностью данной работы является метод представления знаний и метод математического моделирования синтаксических элементов предложений и текстов с использованием математического аппарата и методов искусственного интеллекта. На базе разработанных методов созданы интеллектуальные модули анализа, представления и синтеза текстовой информации на английском языке, которые можно применять для:

- автоматизации перевода текстов с английского языка на другие;
- создания автоматизированной обучающей системы;
- смыслового анализа текстовой информации;
- автоматической генерации модели изучаемого текста;
- автоматического синтеза текстовой информации по полученной модели;
- реферирования и сжатия исходных текстов;
- в задачах обучения – автоматического контроля усвоения материалов пользователем, который обучается английскому языку.

Основная цель системы – организация знаний о лингвистических данных в иностранном языке, процессе его изучения в виде целостной модели и представление этой модели на ЭВМ в доступной для пользователя форме, обеспечивающей изучение этой модели компьютерными средствами через Интернет.

Научная новизна

1. Разработана модель представления и хранения знаний на ограниченной предметной области на английском языке. Особенность модели заключается в объекто – ориентированном подходе к семантическому анализу и представлению знаний англоязычных текстов.
2. Разработаны математические и интеллектуальные методы работы с моделью. Созданы интеллектуальные программные модули анализа, представления и синтеза текстовой информации, обладающие свойством расширяемости, при различном комбинировании которых можно создать системы различного назначения (обучающие, системы автоматического перевода, диалог с компьютером и т.п.)
3. При решении задач была проведена структуризация и формализация знаний, полученных от эксперта. В качестве способа представления знаний в работе использовано комбинирование способов представления в виде объектно – ориентированных семантических сетей и продукционных правил.