Лекция 7

Тема: Проблемы понимания естественного языка

Проблемы понимания естественного языка, будь то текст или речь, во многом зависят от знания *предметной области*. Понимание языка требует знаний о целях говорящего и о контексте. Необходимо также учитывать недосказанность или иносказательность. Например, даже в таком простом предложении "Ваня встретил Машу на поляне с цветами" нам не понятно, кто же был с цветами: Ваня, Маша или поляна? Еще один пример "Врач бегло говорила *по*-английски". Разбирая это предложение, необходимо в результате разбора зафиксировать, что врач была женщина. Крылатая фраза знаменитого русского лингвиста, академика Л.В.Щербы "Глокая куздра штеко будланула бокра и курдячит бокренка" говорит о том, что такая "непонятная" фраза построена *по* всем правилам русского языка, не вызывает проблем с грамматическим разбором такого предложения, но вызывает проблемы с пониманием. Попробуем сформулировать лишь некоторые *проблемы понимания естественного языка*.

1. Проблема СМЫСЛ-ТЕКСТ. Об этом только что говорилось и приведем еще один пример по этой проблеме. В предложении "Какой завод заказал оборудование для конвертерного цеха в Бельгии?" неясен смысл: был ли сделан заказ в Бельгии или цех находится в Бельгии.
2. *Проблема планирования* возникает при необходимости вести диалог, например, на тему "Куда Вы хотите лететь?". В этом случае нужно глубокое знание предметной области (номера рейсов, время прилета-отлета, цены и т.д.).
3. Проблема равнозначности. Будут ли равнозначны два предложения "У дома стоит слон" и "У дома стоит существо с хоботом и бивнями"? На первый взгляд нет сомнений в равнозначности этих предложений. А если в базе знаний существо с хоботом и бивнями определено двумя значениями: слон и мамонт, то такие сомнения, наверное, появятся.
4. Проблемы моделей участников общения. У участников общения должны быть сопоставимые *модели представления знаний,* необходимая глубина понимания, возможность *логического вывода,* возможность действия.
5. Проблема эллиптических конструкций, то есть опущенных элементов диалога. Например, в пословице "Береги платье снову, а честь - смолоду" вторая часть предложения будет синтаксическим *эллипсисом* (опущен глагол береги).
6. Проблема временных противоречий. Например, в предложении "Я хотел завтра пойти в кино" глагол "хотел" в прошедшей форме сочетается с обстоятельством будущего времени "завтра", что противоречит общепринятой логике.

Закончим с перечислением проблем и поговорим об основных понятиях. В качестве языков для общения с программой могут быть: язык *меню,* язык приказов, анкетный язык. Это регламентированные языки, в них могут работать упрощенные схемы разбора, например, *по* ключевым словам, и эти языки мы не рассматриваем. В качестве *естественного языка* (ЕЯ) мы рассматриваем *подмножество* Ограниченного *Естественного Языка* (ОЕЯ) - это профессионально-ориентированное *подмножество* ЕЯ конечного пользователя. Для разбора ОЕЯ используются программные комплексы, называемые Лингвистическими *Трансляторами* (ЛТ). Возможная структурная схема ЛТ приведена на рис. 1.

Рис. 1. Структурная схема ЛТ

Определим или напомним основные понятия. *Слово* - одна из основных единиц языка, служащая для именования предметов, лиц, процессов, свойств и т.д. Предложение - любое *высказывание*, являющееся сообщением о чем-либо. ***Словосочетание*** - простейшая единица речи, которая образуется на основе подчинительной связи (согласования, управления, примыкания) двух и более слов. *Словосочетание* в отличие от предложения не является, как правило, сообщением. ***Дискурс*** - связный текст. ***Лексема*** - слово во всей совокупности его лексических значений. ***Морфема*** - минимальная законченная часть слова. ***Аффикс*** - прикрепленная к корню часть слова (подразделяется на *префикс*, *суффикс*, *инфикс*). ***Омонимы*** - разные по значению, но одинаковые по написанию слова, *морфемы* и др. единицы языка ("рысь" - бег, "рысь" - животное). ***Синонимы*** - разные по написанию слова, но одинаковые по значению ("орать", "кричать" или "дорога", "путь"). ***Эллипсис*** - опущенные слова в предложении ("Я еду кататься, а ты?"). ***Анафора*** - повторение объектов предложения ("Город пышный, город бедный" - А.С.Пушкин).

Анализ текстов на естественном языке

Как видно из рис.1, разбор текстов на ОЕЯ состоит из четырех этапов.

Морфологический анализ
(*МА*) определяет грамматические признаки для каждой словоформы. Грамматические признаки наиболее важных частей речи приведены в табл. 1.

Таблица 1. Грамматические признаки

Часть речи	**Грамматические признаки**
Существительное	Род, число, падеж, склонение
Прилагательное	Род, число, падеж
Глагол	Время, лицо, число, спряжение, вид
Местоимение	Число, лицо

МА для предложения "На мельнице хранятся разные сорта пшеницы" дает следующие результаты разбора (цифрами обозначен порядок слов в предложении): ((на: предлог, 1)

(мельница: существительное, жен. род, ед. число, предл. падеж, 2) (храниться: глагол, мн. число, наст. время, несовершенный вид, третье лицо, 3) (разный: прилагательное, мн. число, имен. падеж, 4) (сорт: существительное, муж. род, мн. число, имен. падеж, 5) (пшеница: существительное, жен. род, ед. число, родит. падеж, 6))

Таким образом, мы видим, что для ***МА*** необходим словарь основ слов и словоформ с их грамматическими признаками в зависимости от *аффиксов* и окончаний. *МА* состоит из выделения основы и флексий входной словоформы. По основе определяются основные характеристики данной *лексемы*, а по виду флексии определяются грамматические характеристики словоформы по словарю. Как правило, *МА* не вызывает больших трудностей на этом начальном этапе разбора, хотя и является достаточно трудоемким этапом из-за необходимости создания точных словарей.

Синтаксический анализ

(***СА***) определяет синтаксическую структуру входного предложения. Основные правила *синтаксического анализа*, в большинстве случаев, следующие.

Подлежащим в предложении может быть

1. существительное в именительном падеже;
2. местоимение в именительном падеже;
3. имя собственное в именительном падеже.

Сказуемое в предложении - это глагол, связанный с подлежащим и согласованный с ним в лице и числе. Подлежащее и сказуемое, как известно, это главные члены предложения.

Дополнение - это существительное, согласованное со сказуемым в падеже. Прямое дополнение - существительное в винительном падеже ("Я вижу окно"). Косвенное дополнение - дополнение не в винительном падеже, часто с предлогом ("Я ехала домой").

Определение - это прилагательное, связанное с подлежащим или дополнением (связь в роде, числе и падеже - это сильная связь).

Обстоятельство - это, как правило, наречие (неизменяемая часть речи - "далеко", "редко") или существительное с предлогом, связанное со сказуемым только семантически.

СА для нашего предложения о пшенице даст следующие результаты: ((На мельнице : обстоятельство места, 1) (хранятся : сказуемое, 2) (разные : определение, 3) (сорта : подлежащее, 4) (пшеницы : дополнение, 5)).

Семантическая интерпретация

(*СИ*) определяет семантическое представление предложения. Результатом *СИ* должна быть модель в виде, например, *семантической сети* (см. лекцию 2) для отображения отношений между объектами предложения или лингвистического фрейма (ЛФ).

Исследование вопросов понимания английского *естественного языка* предложено в работах Хомского (классическая книга по трансформационной грамматике), Вудса , Винограда и других исследователей. Наибольшее распространение в 70-е годы получили расширенные сети переходов (РСП) Вудса и ATNL-грамматики. В английском языке проблема *СИ* упрощается за счет фиксированного порядка слов в предложении.

Например, предложение на английском языке: "The dog has bitten John" переводится как "Собака укусила Джона" или "Джона укусила собака". В русском языке любой вариант перевода правильный и допустим. В английском языке возможен только единственный вариант построения такой фразы:

подлежащее (The dog) => сказуемое (has bitten) =>
=> дополнение (John).

Таким образом, использованию РСП и ATNL-грамматик для разбора русского ОЕЯ в чистом виде препятствуют нефиксированный порядок слов в предложениях русского языка, а также синтаксическая неоднозначность грамматических категорий в предложении. Эти ограничения на структуру фраз русского языка делают метод РСП малоэффективным.

Для *СИ* может быть использован метод семантических падежей К. Филмора , получивший развитие в его работе для разбора русского ОЕЯ. Рассмотрим этот метод подробнее. Предложения выражают чаще всего действия, которые будем отображать в виде предиката в модели на основе ЛФ. Под предикатом в данном случае понимается любой элемент или группа элементов, выполняющих функции сказуемого в предложении, а также *атрибутивные* формы глагола - причастие, деепричастие, инфинитив. Предикат имеет признаки (модальность, переходность, время, наклонение, возвратность, безличность и т.д.), которые являются необходимыми компонентами для правильной *семантической интерпретации* остальных членов предложения из внешней (грамматической) во внутреннюю (семантическую) структуру. Остальные члены предложения разбиваются на группы сильносвязанных слов, в которых выделяется главное слово (как правило, существительное). В группу его актантов включаются причастия, прилагательные, числительные, местоимения, неопределенно-количественные слова и т.д. Главные слова группы являются актантами предиката и выполняют различные семантические "роли", которые можно описать на основе семантических падежей К. Филмора: агент, объект, цель и т.д., а также дополнительные падежи: адресат, добавочный предикат, инструмент, время, место, *определитель*, указатель, количество, пример, деталь и т.п.

Целью *СИ* является однозначное выражение смысла предложения в известных системе внутренних понятиях, отношениях и фактах, а также выделение понятий "новой" декларативной информации, приказа для повелительных предложений и вопросительного элемента для вопросительных предложений. *СИ* включает в себя следующие этапы.

1. Грамматическое и семантическое соотнесение очередного анализируемого элемента с уже разобранными элементами. Объединение элементов в группы сильносвязанных слов с проведением проверки "тестов ожидания" аналогично РСП. С помощью "тестов ожидания" можно проверить наличие фиксированных синтаксических конструкций, информация о которых хранится во входном словаре. Бинарная таблица отношений содержит пары определяемого и зависимого лексических элементов с указанием их грамматико-семантических признаков и семантической роли зависимого слова.
2. Завершение оформления элементов в группы сильносвязанных слов с выделением главного слова, определением семантических ролей внутри группы и определением общих грамматических признаков группы (род, число, падеж и т.д.) на основе информации из словаря. Главное слово в группе выделяется с помощью фильтров модуля СУЩЕСТВИТЕЛЬНОЕ.

3. Определение предиката и его признаков по словарю и выделение в случае группы предикатов главного, связки, глагола "быть", предикативных элементов. Форма предиката (простая, составная глагольная, составная именная) выделяется с помощью фильтров модулей ПРЕДИКАТ. По грамматико-семантическим признакам предикаты разбиваются на несколько КЛАССОВ, указанных в словаре, которые необходимы для выбора формы предиката. КЛАССАМИ предикатов могут быть: ДЕЙСТВИЕ, ФАЗА, СОСТОЯНИЕ, ОБЛАДАНИЕ, РАСПОЛОЖЕНИЕ, ПЕРЕМЕЩЕНИЕ, МОДАЛЬНОСТЬ, КУПЛЯ, ПРОДАЖА, ОТВЛЕЧЕННОЕ ДЕЙСТВИЕ.
4. По окончании входной последовательности слов производится выбор ЛФ-шаблона по классу и типу предиката. В зависимости от типа (личные, безличные, страдательный залог) выбирается соответствующая модификация шаблона. Осуществляется заполнение ЛФ-шаблона с помощью таблиц *бинарных отношений* предикатов и существительных. В случае неопределенности происходит выделение дополнительных связей между группами существительного с помощью этих же бинарных таблиц. В случае неоднозначности связей используется ряд эвристических правил (принцип близости, принцип приоритетности предиката и т.д.) или обращение в базу знаний ИС.

Завершая описание этапа *СИ*, приведем результат *семантической интерпретации* нашего предложения "На мельнице хранятся разные сорта пшеницы" в виде *семантической сети*, показанной на рис. 2.

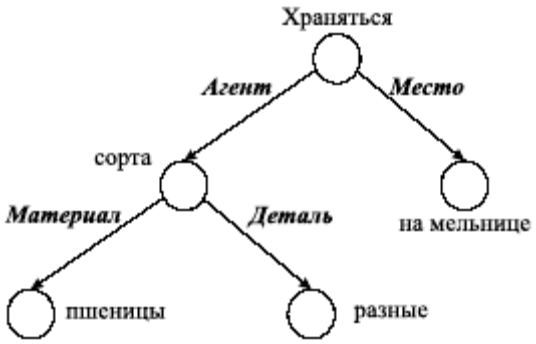

Рис. 2. Результат семантической интерпретации предложения

В виде лингвистического фрейма это выглядит следующим образом:

```
(предикат (хранятся)
   (агент (сорта)
      (материал (пшеницы))
      (деталь (разные)))
   (место (на мельнице)))
```

В этом примере в группе сильносвязанных слов ("разные сорта пшеницы") выделены свои семантические отношения "материал" и "деталь" у главного слова в группе ("сорта").

Проблемный анализ

На этом этапе осуществляется отображение входной строки, представленной в виде сети ЛФ, в сеть проблемных фреймов (ПФ), служащую для внутреннего *представления знаний* в ИС. Для каждой предметной области должна быть разработана собственная база знаний, включающая абстрактную сеть ПФ. "Верхние уровни" ПФ фиксированные и содержат

факты, всегда истинные в предполагаемой ситуации. Нижние уровни содержат много терминалов, то есть "ячеек", которые надо заполнить конкретными данными. В каждом терминале могут перечисляться условия, которым такие присваивания значений обязаны удовлетворять. Например, при анализе зрительных сцен различные фреймы в системе соответствуют различным точкам зрения на нее, а переходы от одного фрейма к другому отражают эффект перемещения наблюдателя из одного места в другое. Важно, что различные фреймы одной и той же системы, будь то ЛФ или ПФ, используют общее множество терминалов. Благодаря этому становится возможным координировать информацию, полученную с различных точек зрения в широком смысле.

Однако именно на этом этапе возникает множество вопросов с пониманием смысла и пониманием причинно-следственных связей. Рассмотрим лишь несколько фраз и связанных с ними вопросов, отмеченных в книге П.Уинстона:

"Робби нравится ставить эту *пирамиду* на красный блок".

Это то же самое, что и фраза "Робби счастлив, когда эту *пирамиду* помещают на красный блок."? В таком случае имеет место

(фрейм ИЗМЕНЕНИЕ СОСТОЯНИЯ
(объект РОББИ)
(текущее состояние ())
(результирующее состояние (БОЛЕЕ СЧАСТЛИВ))
(действие (ФРЕЙМ ДЕЙСТВИЕ)))

Предположим, что кто-то сказал:

"Робби успокоил Суззи".

Ясно, что здесь также присутствует изменение состояния (Суззи стала менее грустной). Но что именно сделал Робби? Поговорил ли он с ней, взял на прогулку или просто передвинул *пирамиду*? Это неизвестно, и поэтому в слоте "действие" ссылка будет отсутствовать.

Рассмотрим еще один пример:

"Суззи была травмирована результатом экзамена."

Ясно, что это образное выражение, экзамен не повредил Суззи физически, а полученная оценка заставила ее почувствовать себя плохо. Приведенный пример снова показывает, что глагол в предложении может указывать на изменение состояния и не отражать действия.

Все сказанное пока о *проблемном анализе* - это рассуждения о моделях на уровне *здравого смысла*. Именно здесь проблемы *естественного языка* попадают в область проблем мышления и понимания и происходит обращение к большим базам знаний. Именно здесь предстоит еще решить множество трудных задач.

Конкретный алгоритм заполнения сети ПФ на основе сети ЛФ читателю предлагается придумать самостоятельно. При этом предлагается подумать также над следующими вопросами:

- Как производится сопоставление фреймов, когда имеются некоторые различия при рассмотрении вещей с различных точек зрения в широком смысле?
- Сколько фреймов нужно для того, чтобы справиться с различными предметными областями, такими, как мир кубиков, мир финансов, политики и окружающий нас мир?
- Как можно усвоить новые фреймы через обобщение старых? Как аналогия может связать различные миры так, чтобы новые фреймы для одного из них можно было бы автоматически строить по фреймам другого?

Завершая наше краткое рассмотрение вопросов общения с ЭВМ на *естественном языке* следует назвать в качестве примеров ранних эффективных отечественных систем для ОЕЯ-общения системы ПОЭТ и АДАЛИТ . В области лингвистической ЕЯ-обработки за последний десяток лет также имеются несомненные успехи : появились коммерческие системы машинного перевода (*Stylus,* Socrat, Pars, Lingvo и др.), поиска информации в ЕЯ-текстах и аннотирования ("Следопыт", "Либретто") и др., создан широкий спектр экспериментальных систем обработки ЕЯ-текстов. Г. Хахалин отмечает также задачи, требующие дальнейшей проработки: трансляция связных ЕЯ-текстов в пределах абзацев и более; полноценный лингвистический синтез текста; автоматизация процесса наполнения моделей; методы проверки ЛТ и *лингвистических моделей* на полноту, корректность и разнообразие. Следует также отметить недостаточную проработанность вопросов унификации моделей проблемной среды, механизмов вывода для ЛТ.

Системы речевого общения

В системах ЕЯ-общения обычно предполагается, что в качестве средства общения используется текст или письменная речь. Поэтому в системах ЕЯ-общения под текстом понимается орфографический текст (как пишется), а в системах речевого общения (СРО) используется фонемный текст (как слышится). В СРО решаются задачи преобразования "текст - речевой сигнал" (синтезатор речи) и "речевой сигнал - текст" (анализатор речи). Синтез речи - это возможность обработки текстовой или числовой информации, согласно установленным правилам произношения для конкретного языка, и преобразование ее в синтезированный голос, *по*восприятию близкий к человеческому. *Анализ* речи - это *распознавание* отдельных слов или слитной человеческой речи, с последующим ее преобразованием в текст либо последовательность команд.

На рис. 3 показано общее *место* анализатора и синтезатора речевых сообщений в *потоке информации*.

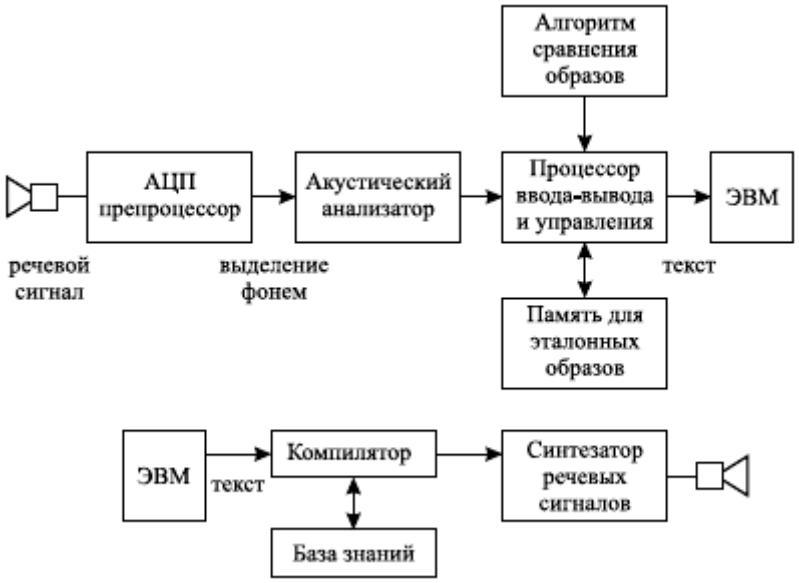

Рис. 3. Анализатор и синтезатор речевых сообщений в потоке информации

Первые СРО стали появляться в конце 70-х годов. Это было связано со следующими преимуществами СРО:

1. удобство, простота и естественность процедуры общения, требующей минимума специальной подготовки;
2. возможность использования для связи с ЭВМ обычных телефонных аппаратов и *телефонных сетей*;
3. устранение ручных манипуляций с одновременным увеличением скорости ввода информации (в 3--5 раз быстрее по сравнению с клавиатурным вводом) и разгрузка зрения при получении информации. Первое и второе преимущество с наибольшим эффектом стали находить применение в *автоматизированных системах управления* (*АСУ*). Третье свойство весьма эффективно может применяться при создании систем оперативного человеко-машинного управления сложными объектами (управление движением, энергетическими установками и т.д.).

Обучающие системы, *синхронный* перевод с одного языка на другой, говорящие книжки, говорящие компьютеры для слепых, управление голосом инвалидными колясками, приборы для генерации и восприятия речи глухонемыми - вот лишь неполный перечень применения СРО.

В основе СРО лежит работа с фонемами. Фонема - это минимальная смысловая *единица* речи. В русском языке 42 фонемы: 6 гласных и 36 согласных. В английском языке 20 гласных (из них 5 дифтонгов) и 24 согласных, во французском - 16 гласных и 20 согласных.

Акустические характеристики фонем обусловлены местом и способом их образования. *По* месту образования фонемы делятся на губные (п, б, ф, в, у, м), зубные и межзубные (д, о), альвеолярные (с, з, р, а), заальвеолярные (ш, ж, щ, э), небные (к, г, х, и, ы) и фарингальные (гортанный, например, английское h). *По*способу образования фонемы делятся на взрывные (п, б, д), аффрикаты (ц, и), щелевые (ф, с, х, в, з, ш, ж,…), дрожащие (р), носовые (м, н), боковые (л), плавные (й), гласные (у, о, а, э, и, ы).

В потоке речи характеристики фонем меняются, что приводит к появлению у них оттенков - аллофонов , например, огубление согласных перед гласными, а также это обусловлено положением фонемы *по* отношению к ударному слогу, концу и началу слова и т.д.

Интонация и ударение определяют направленность высказывания, *логический* смысл, выделение главного и общего (рема и тема), вычленение семантически связанных отрезков речи. Интонация и ударение определяют просодию речи с помощью следующих акустических средств:

- мелодика - изменение частоты основного тона голоса;
- ритмика - текущее изменение длительности звуков и пауз;
- энергетика - текущее изменение интенсивности звука.