ОСНОВНЫЕ НАПРАВЛЕНИЯ КОМПЬТЕРНОЙ ЛИНГВИСТИКИ

Мансурова Тахмина Тахировна
Старший преподаватель
Чирчикский государственный педагогический университет

Аннотация: *В статье рассматриваются основные принципы компьютерного анализа текстов на естественном языке. Приведены примеры анализа на трех уровнях с использованием соответствующего инструментария: словарей и корпусов текстов.*

Ключевые слова*: естественном языке, Джорджтаунский проект, мини-текст, синтаксический, семантический, прагматический, машинный перевод, автоматический перевод, понимание текстов.*

КОМПЬЮТЕР ЛИНГВИСТИКИНИНГ АСОСИЙ ЙЎНАЛИШЛАРИ

Мансурова Тахмина Тахировна
Катта ўқитувчи
Чирчиқ давлат педагогика университети

Аннотация: *Мақолада табиий тил доирасида матнни компьютер таҳлилининг асосий тамойиллари яратилган. Уч даражада таҳлил намуналар и тегишли ҳолатлар қўлланилган ҳолда келтирилган.ҳолатлар қўлланилган ҳолда келтирилган.*

Калит сўзлар*: табиий тил, жорҗтуан лойиҳаси, минитекст, синтактик, семантик, прагматик, машина таржимаси, автомат таржима, матнни тушуниш.*

MAIN DIRECTIONS OF COMPUTER LINGUISTICS

Mansurova Takhmina Takhirovna
Senior Lecturer
Chirchik State Pedagogical University

Abstract*: The article deals with the basic principles of computer analysis of natural language texts. Examples of analysis at three levels using the appropriate tools: dictionaries and text corpora are given.*

Keywords*: natural language, Georgetown Project, minitext, syntactic, semantic, pragmatic, machine translation, automatic translation, text comprehension.*

Введение

Так как вопросов, изучаемых компьютерной лингвистикой, немало, то со временем в ней выделился ряд направлений, посвященных отдельным аспектам автоматической обработки естественного языка. В настоящее время в компьютерной лингвистике выделяют пять основных направлений (Информатика).

1. Анализ текстов на естественном языке. Лингвисты давно изучают, как устроен текст, и прежде всего предложение, играющее роль кирпичика, из совокупности которых складывается текст. Но лишь с появлением компьютеров эти исследования приобрели новое направление. Группа американских лингвистов выдвинула дерзкую идею, получившую название Джорджтаунский проект, — автоматизировать процесс перевода

текстов с одного языка на другой, используя для этого ЭВМ. Идея заинтересовала лингвистов многих стран и активизировала работы в области анализа текстов.

В ходе этих работ надо было ответить, прежде всего, на вопрос: "Существуют ли строгие формальные правила, по которым строится структура предложения и структура текста?" Если о структуре предложения лингвисты накопили много материала, то структура текста ими не изучалась.

В результате проведенных исследований стало ясно, что за каждым текстом (в том числе и за отдельным предложением, являющимся своего рода мини-текстом) скрывается не одна, а несколько формальных структур, которые можно разделить на три уровня (Информатика)

- синтаксический
- семантический
- прагматический.

Более подробно эти и другие уровни анализа текстов естественного языка будут рассмотрены ниже.

Как указывалось выше, направление анализа текстов на естественном языке появилось в связи с желанием решить проблему машинного перевода. Машинный перевод — это автоматический перевод текстов с одного языка на другой (например, пословный перевод научно-технической информации, патентов, документов, инструкций, программ ЭВМ с алгоритмического на машинный язык), а также научное направление, охватывающее круг проблем, которые возникают при автоматизации перевода. Система машинного перевода обычно содержит лингвистические описания входного и выходного языков, т.е. языков исходного текста и текста, полученного в результате перевода, и алгоритм, на основе которого выполняется данный перевод (Информатика).

Со временем (в 50-х гг. 20-го в.) проблема машинного перевода переросла в отдельную научно-техническую проблему и фактически обрела черты отдельного научного направления с одноименным названием. Это направление возникло на стыке таких наук, как математика, кибернетика, лингвистика и программирование. Тем не менее, основу машинного перевода как научного направления составляют результаты, полученные в области компьютерной лингвистики.

2. Синтез текстов на естественном языке. Задача синтеза может рассматриваться как обратная по отношению к анализу. Если заданы некоторая тема и цель будущего текста, то можно считать заданной прагматическую структуру текста. Ее надо декомпозировать в прагматические структуры отдельных предложений и для каждого предложения пройти все этапы анализа в обратном направлении. На сегодняшний день здесь еще масса нерешенных проблем. Неизвестно, как генерировать прагматическую структуру текста из тех целей, которые стимулируют создание текста. Непонятно, как эту структуру разбить на прагматические структуры предложений и как от этих частных прагматических структур перейти к глубинным семантическим структурам. Более известны методы дальнейшего движения по пути генерации текста.

Одним из первых примеров естественно-языковых систем, способных синтезировать тексты, является автоматическая система создания текстов волшебных сказок, созданная в Московском энергетическом институте в 70-х гг. и называемая TALE (Информатика). На первом шаге она выдает тексты примерно такого вида: "Жил-был Х. Не было у него

желаемого У. Стал просить Х Бога. Бог обещал. Появился У. Вырос У. Ушел раз Х и не велел У делать Z. Но У сделал Z. Вернулся Х. У нет. Понял Х, что У сделал Z. Пошел Х искать У..." В памяти рассматриваемой системы хранились данные для заполнения так называемых актантов, а одинаковые переменные показывают, что на эти места всюду надо поставить одни и те же заполнители. Так возникает текст: "Жил-был царь. И не было у царя желаемого наследника. Стал царь просить Бога. Бог обещал. Появился наследник. Вырос наследник..." Вот пример сказки, сочиненной этой программой.

ОДНАЖДЫ В ТРИДЕВЯТОМ ЦАРСТВЕ, В ТРИДЕСЯТОМ ГОСУДАРСТВЕ ЖИЛ ЦАРЬ.

ЦАРЬ ИМЕЛ ДОЧЬ.

ЦАРЬ ОТПРАВИЛСЯ НА ОХОТУ ПООХОТИТЬСЯ.

ЦАРЬ ЗАПРЕТИЛ ДОЧЕРИ ВЫХОДИТЬ ИЗ ДОМА.

НАРУШИЛА ДОЧЬ ЗАПРЕТ. НАЛЕТЕЛ КОЩЕЙ.

УНЕС КОЩЕЙ ДОЧЬ.

ИВАН ОТПРАВИЛСЯ КУДА ГЛАЗА ГЛЯДЯТ ИСКАТЬ ДОЧЬ.

ДОЛГО ЛИ, КОРОТКО ЛИ ШЕЛ ИВАН. ВСТРЕТИЛ ИВАН СТАРУШКУ.

СТАРУШКА ПОГИБАЛА, УМИРАЛА С ГОЛОДА. ПОМОГ ИВАН СТАРУШКЕ, НАКОРМИЛ.

РАССКАЗАЛ ИВАН СТАРУШКЕ, КУДА ПУТЬ ДЕРЖИТ.

ДАЛА СТАРУШКА ИВАНУ КЛУБОЧЕК, КУДА ПОКАТИТСЯ, ТУДА И ИДИ.

ПОКАТИЛ ИВАН КЛУБОЧЕК. ПОШЕЛ ДАЛЬШЕ ИВАН. ВСТРЕТИЛ ИВАН СТАРЕНЬКУЮ СТАРУШКУ. СТАРЕНЬКАЯ СТАРУШКА ПОГИБАЛА БЕЗ ВОДЫ. ПОМОГ ИВАН СТАРЕНЬКОЙ СТАРУШКЕ, НАПОИЛ.

РАССКАЗАЛ ИВАН СТАРЕНЬКОЙ СТАРУШКЕ, КУДА ПУТЬ ДЕРЖИТ.

ДАЛА СТАРЕНЬКАЯ СТАРУШКА ИВАНУ СЕРЕБРЯНЫЙ КЛУБОЧЕК, КУДА ПОКАТИТСЯ, ТУДА И СТУПАЙ СЕБЕ. ПОКАТИЛ ИВАН СЕРЕБРЯНЫЙ КЛУБОЧЕК. ПОШЕЛ ДАЛЬШЕ ИВАН.

ВСТРЕТИЛ ИВАН СОВСЕМ СТАРЕНЬКУЮ СТАРУШКУ.

СОВСЕМ СТАРЕНЬКАЯ СТАРУШКА ПОГИБАЛА, ПАДАЛА ПОД ТЯЖЕСТЬЮ НОШИ. ПОМОГ ИВАН СОВСЕМ СТАРЕНЬКОЙ СТАРУШКЕ ДОНЕСТИ НОШУ.

РАССКАЗАЛ ИВАН СОВСЕМ СТАРЕНЬКОЙ СТАРУШКЕ, КУДА ПУТЬ ДЕРЖИТ. ДАЛА СОВСЕМ СТАРЕНЬКАЯ СТАРУШКА ИВАНУ ЗОЛОТОЙ КЛУБОЧЕК, КЛУБОЧЕК ПОКАТИТСЯ, А ТЫ ЗА НИМ ИДИ.

ПОКАТИЛ ИВАН ЗОЛОТОЙ КЛУБОЧЕК. ПОШЕЛ ДАЛЬШЕ ИВАН.

ПРИШЕЛ ОН В ПОДЗЕМНОЕ ЦАРСТВО КОЩЕЯ.

ВИДИТ ИВАН ЗАМОК ИЗ ЗОЛОТА И СЕРЕБРА. ВОШЕЛ ИВАН В ЗАМОК. СИДИТ В ЗАМКЕ КОЩЕЙ БЕССМЕРТНЫЙ. СПРАШИВАЕТ КОЩЕЙ ИВАНА: ЗАЧЕМ ПОЖАЛОВАЛ КО МНЕ?

ОТВЕЧАЕТ ИВАН КОЩЕЮ: ИЩУ ДОЧКУ ЦАРСКУЮ, ЧТО ТЫ УКРАЛ.

ГОВОРИТ КОЩЕЙ: ВЫПОЛНИШЬ РАБОТУ, ЧТО Я ЗАДАМ, — ТВОЯ ЦАРЕВНА, НЕ ВЫПОЛНИШЬ — ДО КОНЦА ЖИЗНИ ПОД ЗЕМЛЕЙ ОСТАНЕШЬСЯ. КОЩЕЙ ЗАДАЛ ИВАНУ РАБОТУ: ЗА ОДНУ НОЧЬ ВЫРУБИТЬ ДРЕМУЧИЙ ЛЕС, ЗЕМЛЮ ВСПАХАТЬ, ПШЕНИЦУ ПОСЕЯТЬ, МУКУ СМОЛОТЬ, ПИРОГОВ НАПЕЧЬ И МНЕ НА СТОЛ ПОДАТЬ!

ИВАН ВЫПОЛНИЛ РАБОТУ, ПРИНЕС ПИРОГИ. КОЩЕЙ ЗАДАЛ ИВАНУ РАБОТУ: ЗА ОДНУ НОЧЬ ПЧЕЛ РАЗВЕСТИ, ВОСК СОБРАТЬ, ДА ИЗ ВОСКА ДВОРЕЦ ПОСТРОИТЬ.

ИВАН ВЫПОЛНИЛ РАБОТУ. К УТРУ-СВЕТУ БЫЛ ГОТОВ ДВОРЕЦ ИЗ ВОСКА.

КОЩЕЙ ЗАДАЛ ИВАНУ РАБОТУ: ПРИЙТИ НА ЗЕЛЕНЫЙ ЛУГ, ПОЙМАТЬ ТАМ КОНЯ НЕЕЗЖЕНОГО, ДА ПРИЕХАТЬ КО МНЕ НА ТОМ КОНЕ!

ИВАН ВЫПОЛНИЛ РАБОТУ, ОБЪЕЗДИЛ КОНЯ. КОНЬ ШАТАЕТСЯ, ИЗО РТА ПЕНА ПАДАЕТ. ОСВОБОДИЛ ИВАН ЦАРЕВНУ.

ВЗЯЛ ИВАН ЦАРЕВНУ. ПОВЕЗ ИВАН ЦАРЕВНУ ВО ДВОРЕЦ. ЖЕНИЛСЯ ИВАН НА ЦАРЕВНЕ. ИВАН ПОЛУЧИЛ ПОЛЦАРСТВА.

В качестве другого примера системы автоматического синтеза можно привести систему, умеющую слагать стихи. Примеров таких систем уже немало, хотя с сочинением стихов дело обстоит несколько сложнее. Программа в состоянии учесть характер рифмы, количество слогов в соответствующих строках стихотворения, метрическую форму стихотворения и провести грамматический анализ. Однако содержание будущего стихотворения в значительной степени оказывается произвольным и зависит от содержания исходного словаря. Вот пример машинного стихотворения [XXX] (Информатика).

Добрый реет шелест
Плачет пустота
Слушают качели
И поет беда
Стань покорно горе
Томно лишь летит
И прозрачно море
Тайно шелестит
И бежит земная
Незаметно тень
Медленно лесная
Славит влажный день.

В настоящее время автоматический синтез текстов необходим также при создании:

- интеллектуальных систем, способных объяснять пользователю ход решения той или иной задачи,
- систем поддержки принятия решений, способных помогать пользователю принять то или иное решение на основе выработанных альтернатив,
- информационно-справочных систем различного назначения, способных подсказать пользователю наиболее оптимальный путь поиска и/или заказа той или иной информации (например, в системах заказа билетов),
- различных диалоговых систем (см. ниже).

3. Понимание текстов. Проблема понимания текстов на естественном языке включает не только лингвистические аспекты. С ней тесно связаны задачи, традиционно решаемые в рамках психологии, философии и семиотики. Рассказывая о проблеме анализа текстов, мы несколько раз ссылались на то, что сам анализ служит инструментом для понимания содержания текста. Пожалуй, самое важное значение проблема понимания имеет в так называемых диалоговых системах.

В начале 70-х гг. специалисты в области искусственного интеллекта удивляли далеких от науки людей системами общения на естественном языке, демонстрирующими на первый взгляд почти безграничные возможности созданного ими интеллектуального интерфейса общения. Любой человек мог общаться с системой на произвольную тему, используя все богатство доступного ему языка. При этом система поддерживала разговор, поражая собеседника своими способностями. Вот один из примеров такого разговора, в котором в качестве интеллектуального интерфейса использовалась программа, реализованная на ЭВМ, получившая название "Элиза" (Информатика).

Собеседник: Сегодня прекрасная погода, не правда ли?

"Элиза": Вы любите прекрасную погоду?

С.: Конечно. Кто же ее не любит?

"Э.": Да. Прекрасная погода — это хорошо.

С.: Интересно, какие виды в этом году на урожай?

"Э.": Вас это интересует?

С.: Ну, более-менее. А вас?

"Э.": Более-менее.

С.: Но все-таки, хотелось бы, чтобы урожай был хорошим.

"Э.": Я вполне согласна с вами.

Внешне диалог напоминает светскую беседу двух не слишком знакомых людей. Это, скорее, игра в общение, чем общение по существу. Но и такая игра занимает в жизни людей немалое место. "Элиза" в подобных разговорах оказывается вполне "на уровне". Многие даже считают, что их просто разыгрывают и с ними общается не программа, реализованная на компьютере, а живой собеседник. Подобные примеры диалоговых систем сейчас можно найти в сети Интернет, где эти системы получили название языковых чатов.

Уровень общения, который демонстрируют "Элиза" и другие подобные ей программы, по сути, самый низкий из всех возможных. Это уровень так называемого фактического диалога. В таком диалоге партнеры (или один из них) практически не слушают друг друга. Это лишь видимость беседы, а для поддержания ее используются стандартные "домашние заготовки".

4. Оживление текста. Это одно из самых молодых направлений в компьютерной лингвистике. Своим появлением оно обязано персональным компьютерам, которые впервые дали возможность организовать общение с пользователем не только путем обмена текстами, но и посредством зрительных образов на экране дисплея.

Одной из особенностей мышления человека (едва ли не основной для возможности самого мышления) является его разномодальность. Психологи пользуются этим термином, чтобы подчеркнуть, что наши представления об окружающем мире и о нас самих могут иметь различную природу (различную модальность). Можно "мыслить словами", но можно представлять себе какие-то зрительные картинки, как часто бывает в снах. Есть люди, для которых многие воспоминания состоят из запахов или вкусовых впечатлений. Словом, все наши органы чувств дают свою модальность в мышлении. Но две модальности: символьная (текстовая) и зрительная — являются для человека основными (Информатика).

Легко проверить, что между этими модальностями имеется весьма тесная связь. Обычно называние чего-то или текстовое описание некоторой ситуации тут же вызывает зрительные представления об этих объектах и ситуациях. И наоборот, стоит нам увидеть

нечто, как мы тут же готовы описать увиденное с помощью нашего родного языка. Так текст и сопутствующая ему зрительная картина оказываются объединенными в нашем сознании и интегрированными в некоторое единство. Текст как бы "живет" в виде некоторого образного представления. И изучение того, как происходит эта интеграция и как по одной составляющей представления появляется вторая, — одна из увлекательных задач, стоящих перед специалистами в области компьютерной лингвистики и их коллегами — создателями интеллектуальных систем. Уже найдены некоторые важные законы интеграции текстов и зрительных образов. Созданы первые экспериментальные модели этого процесса и первые интеллектуальные системы, способные описывать в виде текста предъявляемую им картинку (например, пейзаж), а также воссоздавать одну из возможных картин, соответствующих введенному в систему тексту.

5. Модели коммуникации. Появление искусственных систем, способных воспринимать и понимать человеческую речь (пока в весьма ограниченном объеме) и тексты на естественном языке, создало предпосылки для непосредственного общения человека и компьютера. Это, в свою очередь, повысило интерес лингвистов к процессам, сопутствующим организации и ведению диалога. Примерами могут служить:

- способ построения сценария диалога на основе тех целей, которые активная сторона в диалоге ставит перед собой;
- поддержка выбранного сценария с учетом интересов партнера и его возможного противодействия тому сценарию, который используется;
- нахождение средств маскировки истинных намерений говорящего;
- организация пассивной поддержки коммуникационного процесса и т.д.

Эти пять направлений, которые активно развиваются в компьютерной лингвистике, естественно, не исчерпывают всего содержания этой науки. Но и сказанного вполне достаточно, чтобы оценить ее важность и значимость не только для самой лингвистики, но и для создания технических систем, по способностям к диалогу, не уступающих человеку.

Литература

1. Апресян Ю.Д. Избранные труды, том I. Лексическая семантика: 2-е изд., испр. И доп. – М.: Школа «Языки русской культуры», Издательская фирма «Восточная литература» РАН, 1995
2. Апресян Ю.Д. Избранные труды, том II. Интегральное описание языка и системная лексикография. – М.: Школа «Языки русской культуры», 2005.
3. Попов Э.В. Общение с ЭВМ на естественном языке. М. Наука. 2000.