ПРИНЦИПЫ РАБОТЫ СИСТЕМ МАШИННОГО ПЕРЕВОДА

Колесник А.С.
Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: kolesnik_nastya23@mail.ru

Машинный перевод – это выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

Современный машинный, или автоматический перевод осуществляется с помощью человека: пред-редактора, который тем или иным образом предварительно обрабатывает подлежащий переводу текст, интер-редактора, который участвует в процессе перевода, или пост-редактора, который исправляет ошибки и недочеты в переведенном машиной тексте.

Для осуществления машинного перевода в компьютер вводится специальная программа, которая реализует алгоритм перевода, под которым понимается последовательность однозначно и строго определенных действий над текстом для нахождения переводных соответствий в данной паре языков при заданном направлении перевода (из одного конкретного языка в другой).

Извлечение информации из текста производится на основании набора атрибутов: морфологических, синтаксических, лексических, семантических и т.п. Атрибуты не указаны в тексте в явном виде, их нужно предварительно получить. Для этого производятся различные виды анализа текста с целью выделения атрибутов, используемых алгоритмом извлечения информации. Анализ, как правило, носит многоуровневый характер и выполняется модулем лингвистического процессора. Обычно выделяют следующие составляющие анализа текста:

- графемный анализ (выделение слов и предложений);
- морфологический анализ;
- синтаксический анализ;
- семантический анализ;
- построение модели предметной области (сценария или ситуации).

На каждом уровне фрагментам текста сопоставляются новые атрибуты. На основании таких наборов атрибутов алгоритм извлечения информации выполняет поиск фрагментов текста, релевантных цели. Естественно, не всегда нужно использовать все уровни текста в полном объеме. Все зависит от предметной области, информации, которую нужно извлечь, источников информации, а также точности и полноты, с которой эту информацию нужно извлекать.

Графематический анализ - это программа начального анализа естественного текста, вырабатывающая информацию, необходимую для дальнейшей морфологической и синтаксической обработки. В задачу графемного анализа входят:

- разделение входного текста на слова, разделители и т.д.

- сборка слов, написанных в разрядку;
- выделение устойчивых оборотов, не имеющих словоизменительных вариантов;
- выделение фамилии, имени и отчества, когда имя и отчество написаны инициалами;
- выделение электронных адресов и имен файлов;
- выделение предложений из входного текста;
- выделение абзацев, заголовков, примечаний.

Алгоритмы **морфологического анализа** делятся на две группы: словарные и бессловарные. Бессловарные алгоритмы более компактны и производительны, но не обладают высокой скоростью, поэтому их применение целесообразно лишь для выявления простых морфологических атрибутов и только в том случае, если нет требования к высокой точности. Если же предполагается использовать синтаксический анализ, то высокая точность является необходимым требованием, и применяется словарный метод.

Словарный метод предполагает наличие словаря основ и флексий. По словарю отыскиваются допустимые наборы атрибутов для каждой графемы. В случае отсутствия слова в словаре, выполняется предсказание парадигмы (аналогично бессловарным методам). Одной графеме может соответствовать несколько наборов атрибутов. Такие случаи - морфологическая омонимия - довольно часто встречаются в русском языке. Существуют алгоритмы для решения этой проблемы с высокой вероятностью успеха.

Целью **синтаксического анализа** является построение синтаксических групп на одном морфологическом варианте одной клаузы, т.е. одного простого предложения в составе сложного.

Целью компьютерной лингвистики в области синтаксиса является построение автоматизированного анализатора отдельного языка. Этот анализатор должен уметь выделять простые предложения в составе сложного, устанавливать связи между словами и по возможности строить полное синтаксическое дерево предложения.

Чтобы разрешить проблему с анализом синтаксически омонимичных конструкций, необходимо построение дерева синтаксических зависимостей между словами во фразе. В случае удачного разбора предложение сворачивается в полносвязное дерево с единственной корневой вершиной.

Поскольку одна словоформа может соответствовать нескольким грамматическим формам слова, в том числе формам различных слов, в ходе анализа необходимо производить свертку предложения для всех возможных вариантов грамматических форм. Те грамматические формы, которые обеспечивают максимальную свертку дерева (минимальное число висячих вершин), следует считать наиболее достоверными.

Семантический анализ строит семантическую структуру одного предложения. Семантическая структура состоит из семантических узлов и семантических отношений. Семантический узел - это такой объект текстовой семантики, у которого заполнены все валентности, как эксплицитно выраженные в тек-

сте, так и имплицитные - те, которые получаются из экстралингвистических источников. Из определения следует, что семантический узел может быть построен только в самом конце семантического анализа. Собственно говоря, главная цель семантического анализа - построение семантических узлов, которое подразумевает заполнение всех валентностей.

Семантический анализ представляет собой выявление в тексте смысловых связей и групп. Этот тип анализа представляется в виде набора составляющих, направленных на выявление различных семантических связей. Во-первых, это выделение именованных сущностей, объектов, которые имеют различную форму записи в тексте и могут принимать различные значения. Второй полезной составляющей является механизм выявления семантических классов. К семантическому классу относится группа понятий, связанных с одной предметной областью и являющихся одной и той же частью речи. Третий момент связан с расширением кореферентности в тексте. Под кореферентностью понимается ссылка разными словами на один и тот же объект действительности. Четвертым элементом семантического анализа является разрешение анафоры. Анафора - это использование языковых выражений, которые могут быть интерпретированы лишь с учетом другого, как правило, предшествующего фрагмента текста. Разрешение анафоры сводится к установлению связи между анафорическим выражением и его интерпретацией (антецедентом).

Построение модели предметной области

Наиболее сложным, но и приносящим наиболее точные результаты этапом является построение модели ситуации или предметной области, которая описывается в тексте. Этот этап реализует представление в структурном виде, отражающем все значимые смысловые связи, всего текста или набора текстов. Но так как задача построения модели очень сложна, в прикладных системах редко прибегают к ее использованию.

Перевод был важен всегда. Научный прогресс дошел до изобретения машинного перевода, который во многом облегчил жизнь переводчикам. Конечно, и сейчас, существует огромное количество недостатков и в таком, казалось бы, совершенном изобретении. Но мы должны приложить все усилия, чтобы развивать машинный перевод.

Список литературы

1. Сокирко А. Будущее машинного перевода. // Компьютерра. - 2002, №21.

2. Кузнецов П. С., Ляпунов А. А., Реформатский А. А. Основные проблемы машинного перевода. Вопросы языкознания, 1956, № 5.

3. Машинный перевод [Электронная статья]. - [Режим доступа]: http://study-english.info/article065.php