

Лекция 6. Компьютерная терминология и компьютерный перевод

План

- ▶ 1. Компьютерная терминология
- ▶ 1.1. Понятие компьютерной терминологии. Термин как основной объект терминологии.
- ▶ 1.2. Терминологические банки данных.
- ▶ 2. Машинный перевод
- ▶ 2.1 Понятие перевода и машинного перевода.
- ▶ 2.2 Классификация и примеры систем МП.
- ▶ 2.3 Этапы осуществления полностью автоматизированного МП.
- ▶ 2.4 Параметры оценки систем МП.
- ▶ 2.5 Проблемы МП.

1.1. Понятие компьютерной терминологии. Термин как основной объект терминологии.

Одним из перспективных направлений компьютерной лексикографии и прикладной лингвистики в целом является работа над электронными терминологическими словарями и банками данных.

Построением специальных терминологических словарей занимается терминология, представляющая собой особый раздел лексикографии.

В то же время терминология тесно связана с терминоведением - наукой о терминах. Соответственно, компьютерная терминология - это наука о составлении электронных терминологических словарей.

Принципы компьютерной терминологии в общем и целом те же, что и рассмотренные выше принципы компьютерной лексикографии.

Их отличия связаны только с основным объектом словарного описания: в лексикографии это обычное слово или другие языковые единицы (морфема, словосочетание, предложение и т.п.), а в терминологии - термин.

Термин - это слово (словосочетание) метаязыка науки или области практической деятельности человека, имеющее четкое и (по возможности) однозначное определение, требующее специальных знаний из соответствующей профессиональной сферы. Так, слово «Интернет» для обычного человека выступает общеупотребительным, а знакомство с соответствующим понятием ограничивается теми манипуляциями, которые человек производит с Интернетом (выбор провайдера услуг, тарифа, настройка подключения и некоторые другие).

1.2. Терминологические банки данных.

Современные компьютерные технологии позволяют обрабатывать и сохранять большие массивы терминов по различным областям знания.

Такие массивы терминов называются терминологическими базами (банка-ми) данных (ТБД). По количеству задействованных в базе данных языков различаются переводческие (многоязычные) и информационно-нормативные (одно-язычные) ТБД.

Кроме того, термины определенной предметной области собираются и описываются в словарях специальных терминов. Эти словари могут быть дескриптивными и нормативными, общими и частными, толковыми и переводными, алфавитными и тезаурусными.

Большинство электронных терминологических словарей носит дескриптивный характер и представляет термины отдельной отрасли знания. При этом востребованы и толковые (одноязычные), и переводные (двухязычные или многоязычные) словари.

При описании термина важными оказываются следующие его свойства, сопоставимые с отдельными зонами словарной статьи:

- ▶ 1) семантика: связь термина с обозначаемым понятием;
- ▶ 2) словоизменение: особенности образования морфологических форм термина;
- ▶ 3) словообразование: включение термина в словообразовательное гнездо, установление связей между однокоренными словами (ср. прилагательные коммуникативный и коммуникационный, относящиеся к разным значениям термина «коммуникация»);
- ▶ 4) синтаксические связи: управление, сочетаемость с другими терминами и нетерминами;
- ▶ 5) парадигматические связи в терминосистеме: синонимы, антонимы, гиперо-гипонимические связи, пересечения значения, терминологические ряды;
- ▶ 6) произношение;
- ▶ 7) примеры использования в контексте;
- ▶ 8) происхождение;
- ▶ 9) переводные эквиваленты.

2.1 Понятие перевода и машинного перевода.

Вопросы машинного перевода составляют одну из центральных областей использования информационных технологий в лингвистике.

Это обусловлено не только тем, что в машинном переводе как в фокусе концентрируются все проблемы компьютерной лингвистики - от способов анализа содержания до синтеза словоформы, предложения и целого текста, но и постоянно возрастающей практической потребностью современного общества в переводе значительного количества текстов различной функциональной направленности.

Услуги переводчиков обходятся в миллиарды долларов.

Кроме того, что работа переводчика-человека достаточно дорогая, она к тому же весьма медленная. Так, нормой научно-технического перевода считается время 10 дней на авторский лист (24 страницы машинописного текста). Система машинного перевода позволяет получить перевод сотен авторских листов за 1 час.

Кроме того, появляются новые области применения машинного перевода, например, тексты Интернета. По подсчетам исследователей, в Интернете встроенными системами перевода (SYSTRAN, TRADOS и ESTeam Translator) и сетевыми онлайн-словарями ежедневно выполняется 1 млн запросов на перевод текстов в различных форматах.

Все вышесказанное свидетельствует об актуальности обращения к проблеме машинного перевода, который хотя и уступает по качеству переводу, осуществляемому человеком, но даже на сегодняшнем этапе развития позволяет преодолевать языковые барьеры, а кроме того, продолжает оставаться интересной научной проблемой компьютерной лингвистики в целом.

Чтобы определить понятие машинного перевода, обратимся сначала к некоторым общим понятиям теории перевода. Перевод обычно понимается как деятельность, «в результате которой некоторый текст на одном языке ставится в соответствие тексту на другом языке, при этом обеспечивается их смысловая эквивалентность». При этом отмечается многозначность понятия перевода: это одновременно и процесс передачи содержания текста на одном языке средствами другого языка, и результат переводческой деятельности.

2.2 Классификация и примеры систем МП.

Машинный (или автоматический) перевод (МП) - выполняемое компьютером действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

К числу наиболее распространенных в России систем МП относятся:

- ▶ Stylus - система МП, включающая множество словарей по разным предметным областям;
- ▶ Universal Translator— многоязычная система МП;
- ▶ Socrat - система, позволяющая сканировать документы, переводить их содержимое и проверять орфографию;
- ▶ Polyglossum - многоязычная система МП с широким набором предметных словарей;
- ▶ Promt - многоязычная система МП, содержащая множество словарей по разным предметным областям;
- ▶ Web TranSite- система для перевода веб-страниц

С точки зрения роли человека в процессе выполнения МП различают следующие его

ВИДЫ:

- ▶ МАНТ (Machine-assisted human translation) - перевод, осуществляемый человеком с использованием компьютера;
- ▶ НАМТ (Human-assisted machine translation) - машинный перевод при участии человека;
- ▶ ФАМТ (Fully-automated machine translation) - полностью автоматизированный машинный перевод.

- ▶ В первом случае человек использует компьютерные инструменты, направленные на ускорение и упрощение процесса перевода, но собственно перевод текста выполняет сам человек. Вспомогательными системами компьютерной поддержки перевода здесь выступают электронные словари, терминологические базы данных.
- ▶ Второй тип систем МП является своего рода промежуточным: здесь одинаково важно участие в процессе перевода и человека, и машины. В машину вводятся электронные словари, морфологические справочники и задается определенный алгоритм выполнения задачи перевода. Роль человека здесь сводится к выбору предлагаемых машиной решений и редактированию текста перевода.

По мере усложнения систем МП и включения в них новых этапов автоматического анализа и синтеза текста выделяют три поколения таких систем

- ▶ П-системы - системы прямого перевода (direct systems);
- ▶ Т-системы - системы с синтаксическим преобразованием исходного текста (от англ. transfer - преобразование);
- ▶ И-системы - системы с семантическим и прагматическим анализом (interlingua - язык-посредник).

2.3 Этапы осуществления полностью автоматизированного МП.

В целом схема машинного перевода включает следующие этапы:

- ▶ ввод в компьютер текста на ИЯ,
- ▶ его морфологический анализ, т.е. определения части речи и морфологических характеристик каждого слова,
- ▶ синтаксический анализ каждого предложения текста ИЯ (поиск основных членов предложения и определение типов синтаксических связей между ними, выражаемых в виде дерева зависимостей или дерева непосредственных составляющих),
- ▶ семантический анализ каждого предложения ИЯ, в результате которого создается семантическое представление этого предложения, независимое от типа языка (общее и для ИЯ, и для ПЯ),
- ▶ синтаксический синтез предложений ПЯ (создание предложений правильной синтаксической структуры, соответствующей правилам ПЯ и типу синтаксической структуры предложения на ИЯ,
- ▶ морфологический синтез каждого слова в составе отдельных предложений текста ПЯ (постановка слов ПЯ в нужных морфологических формах);
- ▶ вывод текста на ПЯ.

2.4 Параметры оценки систем МП.

- ▶ Сравнение и оценка систем МП осуществляется по следующим параметрам (Framework for the Evaluation of Machine Translation, FEMTI):
- ▶ характеристики программного обеспечения: надежность системы, удобство использования, скорость работы, возможность обновлений, эффективность, мобильность и т.п.;
- ▶ характеристики пользователя и задач перевода: особенности пользователя, автора и текста, а также назначение перевода;
- ▶ особенности системы МП: стратегия построения системы, лингвистические ресурсы и т.п.;
- ▶ специфика выходного текста: точность, целостность, стиль и т.п., а также наличие ошибок любого характера.

2.5 Проблемы МП

Отдельные трудности процесса МП связаны с необходимостью определения анафорических связей в текстовом целом (anaphorare solution), снятия омонимии на разных уровнях, а также с необходимостью привлечения в процесс перевода экстралингвистических знаний.

Необходимость включения
экстралингвистической инф
МП иллюстрируется, к примеру, следующими
фразами:

- ▶ *Председатель Центральной избирательной комиссии назначается президентом Российской Федерации.*
- ▶ *Согласно задумкам американских ученых, сразу после старта вражеские ракеты будут уничтожать авиационные лазеры и мобильные комплексы малых противоракет.*

Лингвистическим обеспечением таких систем выступают словари слов и словосочетаний с соответствующими признаками для ИЯ и ПЯ; морфологические таблицы суффиксов и окончаний для ИЯ и ПЯ; базы грамматических правил и др. К программному обеспечению относятся программы выполнения перевода, ведения словарей, формирования базы правил и т.д. Информационное обеспечение представляет база экстралингвистических знаний о предметной области.