А. Н. Шеремет

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПЕРЕВОДЕ

Использование англоязычных ресурсов наравне с русскоязычными на сегодняшний день не является чем-то сверхъестественным. Наоборот, люди уже привыкли, что всегда под рукой есть тот или иной инструмент, информационная технология, которая даст им возможность получить адекватный перевод любого информационного источника на иностранном языке. К таким удобным и современным инструментам относится лингвистическое программное обеспечение.

Лингвистическое программное обеспечение (Lingware), далее ЛПО - компьютерные программы и данные, обеспечивающие анализ, обработку, хранение и поиск аудиоданных, рисунков и текстов на естественном языке. Существует следующие виды ЛПО:

1. Обработка текста на естественном языке:

    1. Электронные словари: Викисловарь, Мультитран, GoldenDict, Stardict, dict, ForceMem, ABBYY Lingvo и др.;

    2. Орфокорректоры (или спеллчекеры): Орфо, MS Word, ispell, aspell, myspell и др.;

    3. Поисковые системы;

    4. Системы машинного перевода: PROMT, Socrat, Apertium и др. ;

    5. Системы автоматизированного перевода, в том числе программы управления памятью переводов - OmegaT, STAR Transit NXT, Trados, MetaTexis, Deja Vu и др.

2. Системы распознавания символов (optical character recognition, OCR): Finereader, CuneiForm, Tesseract, OCRopus.

3. Речевые системы:

    1. Системы анализа речи: Dragon, IBM via voice;

    2. Системы синтеза речи: Агафон;

    3. Системы голосового перевода (распознавание и синтез): Speereo.

В результате увеличивающегося интереса и возрастающих потребностей в осуществлении перевода больших массивов информации на рынке стали появляться системы машинного перевода.

Машинный перевод, МП (machine translation, MT) - процесс перевода текстов с одного естественного языка на другой с помощью специальной компьютерной программы. Так же называется направление научных исследований, связанных с построением подобных систем. Машинный перевод по-другому можно назвать автоматическим переводом. Это происходит от прямого перевода словосочетания «machine translation» с английского языка на русский.

В тоже время существует такое понятие как автоматизированный перевод, когда используется какое-либо программное обеспечение в качестве помощника при переводе. Эти термины необходимо различать. В английском варианте они выглядят так: machine-aided или machine-assisted translation (MAT) (автоматизированный) и machine translation (MT) (только автоматический перевод). В англоязычных источниках п ри упоминании обоих терминов пишут M(A)T.

По формам взаимодействия с пользователями автоматизированный перевод может быть частично автоматизированным, когда используются компьютерные переводчики и словари в помощь человеку, а может быть - с разделением труда, когда человек дорабатывает за машиной переведенные ею фразы жестко определенной структуры. При этом выделяются следующие виды редактирования:

- постредактирование: сначала текст переводит программа, а затем человек исправляет за ней все недочеты.
- предредактирование: сначала идет подготовка текста к обработке программой (например, человек упрощает текст, убирает словосочетания с двояким смыслом) и только после после этого происходит перевод текста.
- интерредактирование: в процессе перевода программой-переводчиком текста человек вмешивается в работу, помогая разрешить трудные случаи.
- смешанные системы (используется одновременно разные виды редактирования текста).

Системы машинного перевода относятся к классу текстовых процессоров и обладают всеми свойствами этих программ, т.е. позволяют создавать и обрабатывать различные тексты.

В тоже время системы МП являются частью лингвистических процессоров. Лингвистический процессор (далее ЛП) - более сложное понятие, которое относится к такому научному направлению как искусственный интеллект. Лингвистическим процессором называется реализованная на ЭВМ формальная лингвистическая модель, способная понимать и производить тексты на неограниченном языке. Она включает три основных массива правил: морфологические, синтаксические и семантические, а также обслуживающие их словари. Задачей лингвистического процессора является интеллектуальная обработка текстов на естественном языке. Ему необходимо преобразовать этот текст в некоторый набор семантических структур, которые являются формальным представлением «смысла» исходного текста. ЛП выполняет роль посредника между пользователем и базой, хранящей лингвистические данные (Рис. 1).

Рис. 1. Схема взаимодействия пользователя с лингвистической базой данных

Современные требования к системам МП определяются эффективностью этих систем. При оценки их результативности использовались различные методы, начиная со статистических и заканчивая непосредственно учетом мнения пользователей. Тем не менее, на настоящий момент нет единых общепринятых критериев эффективности таких систем. Единственно в чем сходятся специалисты в области машинного перевода - это то, что перевод должен быть максимально понятен и адекватен. Например, существует шкала оценки перевода, предложенная японским ученым Макото Нагао:

1. Смысл предложения понятен и не возникает никаких вопросов. Грамматика, словоупотребление и стиль соответствуют общей структуре текста и не требуют постредактирования.

2. Смысл предложения понятен, но возникают большие проблемы с грамматикой, словоупотреблением и стилем.

3. Общий смысл предложения понятен, но смысл некоторых его частей вызывает сомнение из-за неправильного грамматического строя,

4. Словоупотребления и стилистических ошибок. Требуется обращение к оригиналу.

5. В предложении имеется большое количество грамматических, словоупотребительных и стилистических ошибок. Смысл предложения с трудом можно понять после внимательного изучения.

6. Смысл предложения непонятен.

В проверке и оценке систем МП могут быть использованы различные подходы. Так, например, часто применяется принцип «черного ящика» (black box), когда предположение о внутренней структуре системы МП и о ее типе делается на основании оценки выполненных ею переводов. Происходит это потому что разработчики таких систем стараются скрыть технологию их функционирования и алгоритмы создания перевода. В противовес такому подходу существует принцип прозрачности системы (glass box). Он состоит в поэтапном отслеживании прохождения анализа и синтеза текста, в процессе которого можно определить, какой модуль или алгоритм функционирует неправильно.

Итак, на сегодняшний день в современном машинном переводе существует несколько подходов и технологий. Рассмотрим подробнее каждые из них.

По одной классификации системы МП подразделяются на следующие группы:

- системы прямого типа;
- трансферные системы;
- системы семантического типа.

Системы прямого перевода относятся к процессорам полулексического, полуграмматического типа. Они имеют возможность проводить морфологический анализ и синтез, но у них отсутствуют модули полного синтаксического и семантического анализа и синтеза, что затрудняет воссоздание структуры предложения в терминах формальной грамматики или семантики конечного языка. Из-за этого для таких систем характерно невысокое качество перевода, что неизбежно влечет за собой необходимость постредактирования.

В трансферных системах МП применяется развернутый синтаксический анализ и синтез, более того, анализ осуществляется в категориях входного языка, а синтез в категориях выходного языка. Для того чтобы преобразовать результаты анализа на исходном языке в синтез конечного языка, требуется трансфер (to transfer - переносить, перемещать), поэтому эти системы и называются трансферными, т.е. они используют межъязыковой перенос. Для перевода таких систем характерна большая надежность и адекватность нежели чем систем прямого типа.

Принцип действия систем семантического типа основан на использовании семантических баз знаний. Ключевым отличием этих систем от вышеназванных является наличие компонента, включающего экстралингвистические знания.

Для человека-переводчика важны не только лингвистические знания (иностранный и родной язык), но и экстралингвистические знания (мимика, жестикуляция - в устном переводе, фоновые знания - в письменном переводе). Фоновые знания могут выражаться в знании предметной области, знании культуры другой страны, ее национальные особенности и прочее.

Исходят из этого, говорить о серьезном применении систем МП семантического типа пока раньше времени. Скорее всего это будет возможно только после прорыва в области создания систем искусственного интеллекта. Сначала необходимо научить машину формализовывать и представлять экстралингвистические знания.

Список литературы

1. Википедия - свободная энциклопедия [Электронный ресурс]. - http://wikipedia.org (дата обращения: 26.09.2015).

2. Лингвистический процессор [Электронный ресурс] : лингвистика URL: http://lingold.ru/lingvisticheskij-proczessor/ (дата обращения: 26.09.2015).

3. Научная библиотека диссертаций и авторефератов disserCat [Электронный ресурс]: URL: http://www.dissercat.com/content/analiz-i-otsenka-effektivnosti-sovremennykh-sistem-mashinnogo-perevoda#ixzz3kJBbcDVg (дата обращения: 26.09.2015).

4. Можаров М.С., Шеремет А.Н. Система модульных курсов как содержательная основа для формирования академической мобильности // Философия образования. 2009. № 4. С. 68-73.

5. Сликишина И.В. Некоторые вопросы подготовки бакалавров прикладной информатики в педагогическом вузе // Информация и образование: границы коммуникаций. 2012. Т. 4(12). С. 115-117.

6. Читайло А.И. Возможности применения игрового адаптера ПК в школьном физическом эксперименте // Информационно-коммуникационные технологии в педагогическом образовании. 2009. № 1 (01). С. 20-24.

7. Sato S., Nagao M. Towards memory based translation. In Proceedings of COLING'90, Helsinki, Finland, 1990.