Ю. Б. Крапивин, Д. В. Степанова

ОСОБЕННОСТИ ГИБРИДНОЙ МОДЕЛИ МАШИННОГО ПЕРЕВОДА

Машинный (*автоматический*) *перевод* можно определить как выполняемое компьютером действие по преобразованию текста на одном естественном языке в текст на другом естественном языке при сохранении содержания, а также результат такого действия (Ю. Н. Марчук).

В настоящее время существует 2 основных вида систем машинного перевода: машинный перевод, основанный на правилах (Rule-Based Machine Translation, RBMT), и статистический машинный перевод (Statistical Machine Translation, SMT).

Машинный перевод, основанный на правилах, предполагает анализ лингвистической информации об исходном и переводном языках. Преобразование текста на исходном языке (ИЯ) в текст на переводящем языке (ПЯ) осуществляется постепенно, по предложениям, с опорой на два основных компонента такой системы: лингвистическую базу данных, включающую двуязычные словари, файлы имен, транслитерации, синтаксические, морфологические и семантические закономерности каждого языка, и модуль перевода, состоящий из грамматических правил и алгоритмов перевода.

Принято выделять 3 основные стратегии разработки систем машинного перевода, основанного на правилах:

1) системы пословного перевода (Direct Translation Systems), которые осуществляют относительно неглубокий анализ входного текста без фиксирования его лингвистической структуры и четкого разделения процессов анализа и синтеза. В процессе перевода входной текст преобразуется в текст на ПЯ путем замены всех его элементов, найденных в словаре, на их переводные эквиваленты;

2) трансферные системы машинного перевода (Transfer Systems), которые осуществляют морфологический, лексический и семантико-синтаксический анализ предложения на ИЯ, создают синтактико-семантическое дерево разбора входного предложения, преобразовывают структуры входного предложения в соответствии с формальными требованиями ПЯ и на этапе синтеза формируют конечное предложение на ПЯ;

3) интерлингвистические системы машинного перевода (Interlingua Systems), в основу разработки которых положена теория о том, что любое предложение любого языка может быть преобразовано в его смысловое представление

на так называемом универсальном метаязыке; из полученного смыслового представления можно синтезировать предложение на ПЯ. Иными словами, с помощью определенного набора правил и словаря с семантическими характеристиками можно преобразовывать текст в смысл и наоборот.

Системы машинного перевода, основанного на правилах, характеризуются синтаксической и морфологической точностью результата перевода и возможностью настройки на определенную предметную область. В качестве недостатков такой системы можно отметить трудоемкость разработки ее основных компонентов и необходимость регулярной актуализации лингвистической базы данных.

Статистический машинный перевод представляет собой разновидность машинного перевода, при котором текст на ПЯ генерируется на основе статистических моделей (по словам, фразам, синтаксису, иерархическим фразам), параметры которых являются производными от анализа корпусов параллельных текстов. На первоначальном этапе статистическая система проходит обучение, в процессе которого извлекаются статистические данные о переводе отдельных слов и фраз ИЯ на ПЯ. На основе полученных данных на этапе перевода система вычисляет наиболее вероятный перевод исходного предложения. Данные корпуса текстов система использует при построении статистической модели ПЯ, которая позволяет оценить, насколько вариант перевода соответствует нормам и правилам ПЯ.

Статистические системы машинного перевода отличаются возможностью быстрой настройки и самообучаемости, а также достаточно хорошим качеством результата перевода. Вместе с тем следует отметить, что разработка корпусов параллельных текстов представляет собой отдельную непростую задачу. Кроме того, статистический перевод зачастую содержит грамматические ошибки и в целом характеризуется нестабильностью и непредсказуемостью.

Таким образом, можно утверждать, что основной проблемой машинного перевода остается формализация лингвистических данных от простой морфологической структуры слова до сложной семантической организации целого текста.

Для решения данной проблемы в последнее время особое внимание уделяется разработке гибридной системы машинного перевода, которая содержит 2 основных компонента: базовый модуль перевода, основанного на правилах, и модуль статистического постредактирования, который использует данные, полученные на этапе обучения (статистическая модель перевода, статистическая модель выходного языка). На первом этапе процесса перевода с помощью базового модуля осуществляется преобразование исходного предложения ПЯ; на последующем этапе полученный перевод обрабатывается посредством статистического компонента, т.е. реализуется перевод с формального языка на естественный по правилам статистического машинного перевода.

С опорой на рассмотренные стратегии машинного перевода в Минском государственном лингвистическом университете была разработана автоматизированная информационная система «Англо-белорусский словарь», которая

обеспечивает автоматизацию процесса ведения словарей терминологической лексики по информационным технологиям, а также пословно-оборотного перевода английских текстов выбранной предметной области на белорусский язык в человеко-машинном режиме взаимодействия. Разработанная система состоит из 2 подсистем:

- подсистема ведения словарей терминологической лексики по информационным технологиям;
- подсистема пословно-оборотного перевода английских текстов по информационным технологиям на белорусский язык.

Применение гибридного подхода к решаемой задаче по совершенствованию машинного перевода научно-технических текстов по информационным технологиям с английского языка на белорусский предполагает обязательный контроль по разработанному автоматическому словарю терминологической лексики по информационным технологиям.

Комбинирование рассмотренных выше стратегий машинного перевода обеспечивается с помощью алгоритма анализа результатов, получаемых от их применения в отдельности, основанном на методе обучения CatBoost. При вычислении оценки учитывается ряд факторов, от длины предложения (короткие фразы и слова с невысоким показателем частотности употребления лучше переводятся при использовании статистического подхода) до анализа синтаксической структуры. Алгоритм позволяет оценить оба варианта перевода по всем факторам, выбирая лучший и представляя его в качестве результата.

Реализованная в рамках подсистемы пословно-оборотного перевода английских текстов по информационным технологиям на белорусский язык (модуль обработки запроса / документа, модуль перевода) возможность первоочередного поиска и перевода терминов входного текста на основании терминологических словарей из лингвистической базы данных также позволяет обеспечить:

- автоматический лингвистический анализ входного текста в части сбора сведений о частоте встречаемости словоформ из словарей в тексте, их лемматизации, определения принадлежности к части речи, визуализации дерева грамматического разбора заданного предложения и сохранения собранной информации в соответствующие файлы на жестком диске;
- автоматический перевод текста выбранной предметной области с английского языка на белорусский с учетом информации, содержащейся в разработанных терминологических словарях лингвистической базы данных созданной системы.

Предложенная гибридная модель перевода с контролем по созданным терминологическим словарям доказала высокую эффективность функциональности и может быть использована для формализации процесса перевода терминологической лексики других предметных областей, а также послужить основой для решения лингвистических проблем автоматического перевода.