ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Т. В. Бусел

СОВРЕМЕННЫЕ ПОДХОДЫ К СОЗДАНИЮ СИСТЕМ МАШИННОГО ПЕРЕВОДА

С середины XIX в., когда американским ученым У. Уивером была впервые сформулирована концепция машинного перевода (МП), ведущие научные центры и университеты мира занимаются созданием интеллектуальных компьютерных систем, позволяющих осуществлять «сложные преобразования на всех языковых уровнях двух и более языков для передачи смысловой информации текста одного языка и создания эквивалентного ему по форме и содержанию текста на другом, выходном языке».

Проблема МП довольно сложна и не теряет своей актуальности и в XXI в. по целому ряду причин. Во-первых, спрос на переводы в мире постоянно увеличивается по мере того, как все больше стран приобщается к мировой цивилизации. Перевод с одного языка на другой – единственный эффективный способ обеспечения межъязыковой коммуникации, объем которой возрастает с каждым годом. Другие способы преодоления языковых барьеров на пути коммуникации – разработка или принятие единого языка, а также изучение иностранных языков – не могут сравниться с переводом по эффективности. С этой точки зрения можно утверждать, что альтернативы переводу нет, так что разработка качественных и высокопроизводительных систем МП способствует разрешению важнейших социально-коммуникативных задач.

Во-вторых, высока научная привлекательность проблемы МП, что обусловлено комплексностью и сложностью компьютерного моделирования данного процесса. Как вид языковой деятельности перевод затрагивает все уровни языка – от распознавания графем (и фонем при переводе устной речи) до передачи смысла высказывания и текста. Кроме того, для перевода характерна обратная связь и возможность сразу проверить теоретическую гипотезу об устройстве тех или иных языковых уровней и эффективности предлагаемых алгоритмов. Эта специфическая черта перевода вообще и МП в частности привлекает внимание теоретиков, в результате чего продолжают возникать все новые теории автоматизации перевода и формализации языковых данных и процессов.

При моделировании процесса перевода в автоматизированной системе перевод рассматривается как многоуровневый процесс, где каждая процедура переводит компонент специального уровня. Из этого следует, что исходные конструкции переводимого текста на каждом уровне анализа должны распознаваться, описываться и преобразовываться в выходные конструкции перевода, которые могут быть изменены на следующем уровне в соответ-

ствии с их структурными особенностями. Таким образом, процесс перевода моделируется в системе МП как композиция лексических и семантико-синтаксических процессов.

В зависимости от особенностей морфологии, синтаксиса и семантики конкретной языковой пары, а также направления перевода *общий алгоритм* перевода в системе МП, как правило, включает следующие этапы.

1. На *первом* этапе осуществляется ввод текста и поиск входных словоформ (слов в конкретной грамматической форме, например, родительного падежа единственного числа) во входном словаре (словаре языка, с которого производится перевод) с сопутствующим морфологическим анализом, в ходе которого устанавливается принадлежность данной словоформы к определенной лексеме (слову как единице словаря). В процессе анализа из формы слова могут быть получены также сведения, относящиеся к другим уровням организации языковой системы.

2. *Второй* этап включает:

- перевод идиоматических словосочетаний, фразеологических единств или штампов данной предметной области (например, при англо-русском переводе обороты типа *in case of*, *in accordance with* получают единый цифровой эквивалент и исключаются из дальнейшего грамматического анализа);
- определение основных грамматических (морфологических, синтаксических, семантических и лексических) характеристик элементов входного текста (например, числа существительных, времени глагола, синтаксических функций словоформ в данном тексте и пр.), производимое в рамках входного языка;
- разрешение омографии (конверсионной омонимии словоформ, например, англ. *round* может быть существительным, прилагательным, наречием, глаголом или же предлогом);
- лексический анализ и перевод лексем.

Обычно на этом этапе однозначные слова отделяются от многозначных (имеющих более одного переводного эквивалента в выходном языке), после чего однозначные слова переводятся по спискам эквивалентов, а для перевода многозначных слов используются так называемые контекстологические словари, словарные статьи которых представляют собой алгоритмы запроса к контексту на наличие / отсутствие контекстных определителей значения.

3. На *третьем* этапе происходит окончательный грамматический анализ, в ходе которого определяется необходимая грамматическая информация с учетом данных выходного языка (например, при русских существительных *деньги, фрукты, чернила* глагол должен стоять в форме множественного числа, в то время как в оригинале может быть и единственное число).

4. На *четвертом* этапе осуществляется синтез выходных словоформ и предложения в целом на выходном языке.

Для реализации данного алгоритма перевода в современных системах МП, как правило, используются три типа моделей:

1) статистические модели (*Statistical Machine Translation*, или *SMT*);

2) модели на основе правил (*Rule-Based Machine Translation*, или *RBMT*);

3) гибридные модели (*Hybrid Machine Translation*, или *HMT).*

Для систем *первого* типа характерно использование статистической модели перевода на основе параллельного корпуса обоих языков (содержащей вероятности соответствия слов исходного языка словам языка перевода), а также статистической модели языка на основе корпуса языка перевода (содержащей вероятности следования слов определенному количеству предшествующих слов в данном языке). Данная модель предоставляет возможности улучшить перевод, используя наиболее частотные словоупотребления на различных языках, учитывая в дальнейшем соответствующие частоты при переводе документа. Применение методов статистического перевода позволило успешно решить проблемы снятия смысловой многозначности, разрешения проблемы анафор (например, интерпретация местоимений), сегментации дискурса и др.

Системы *второго* типа производят анализ текста, который используется в процессе перевода. Перевод выполняется на основе встроенных словарей для данной языковой пары, а также грамматик, охватывающих семантические, морфологические, синтаксические закономерности обоих языков. На основе всех этих данных исходный текст последовательно, предложение за предложением, преобразуется в текст на требуемом языке. Главный принцип работы таких систем – связь структур исходного и конечного текстов.

Среди систем МП, основанных на использовании правил, наиболее эффективными являются *трансфертные системы*, которые работают по следующим принципам: проводится морфологический, лексический и семантико-синтаксический анализ предложения на языке оригинала, создается синтактико-семантическое дерево разбора входного предложения, затем производится так называемый «трансфер», т.е. преобразование структуры входного предложения в соответствии с формальными требованиями языка перевода. На заключительном этапе синтеза формируется конечное предложение на языке перевода.

Системы *третьего* типа (гибридные) объединяют технологии машинного перевода на основе правил и на основе статистических моделей, выполняют лингвистический анализ входного предложения, порождение вариантов перевода, использование статистических технологий, оценку и выбор лучшего варианта перевода с использованием модели языка.

Улучшение качества современного МП представляет собой трудоемкую задачу, поскольку перевод – процесс творческий и довольно сложный, для выполнения которого требуется не только хорошая лингвистическая подготовка, но и знание области, к которой относится переводимый текст.