Р. В. Детскина, О. В. Великая

ТЕКСТ КАК ОБЪЕКТ АНАЛИЗА В СИСТЕМАХ АВТОМАТИЧЕСКОГО ПЕРЕВОДА

Бурное развитие научно-технического перевода обусловлено постоянным развитием технической индустрии, международных отношений, подписанием многочисленных договоров и контрактов с зарубежными представителями, увеличением объема коммерческой информации, что в первую очередь отражается в объеме переводимых документов.

Развитие машинного перевода (МП), которое фактически началось в конце 40-х гг. прошлого века, продолжается и по настоящее время. Интерес ученых к машинному переводу не теряет своей актуальности и в XXI веке. Во-первых, спрос на переводы в мире постоянно увеличивается по мере того, как все больше стран приобщаются 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.

Высока также и 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

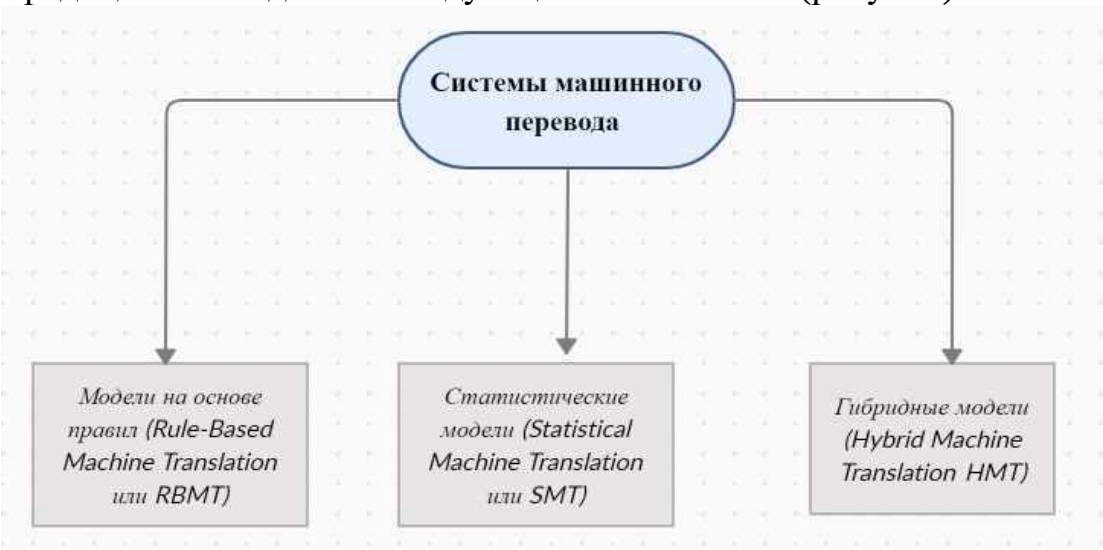
черта перевода в целом и МП в частности привлекает внимание теоретиков, в результате чего продолжают возникать все новые теории автоматизации перевода и формализации языковых данных и процессов.

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 соответствии с их структурными особенностями.

Таким образом, процесс перевода моделируется в системе МП как композиция лексических и семантико-синтаксических процессов. Изначально следует кратко охарактеризовать системы машинного перевода (МП) и раскрыть назначение текста в них.

Традиционно выделяют следующие системы МП (рисунок):



Технология машинного перевода на основе правил использует в своей работе два важных компонента (два словаря – словарь лексики и словарь грамматики): база словарной информации и анализ грамматических правил, охватывающих семантические, морфологические и синтаксические особенности обоих языков. На основе этих составляющих текст последовательно, предложение за предложением, преобразуется в текст на требуемом языке. Основной принцип работы таких систем – связь структур исходного и конечного текста.

Кроме этого, выделяются три подкатегории, которые будут определять то, для каких элементов языковой системы будут подобраны лингвистические правила.

Первая подкатегория – системы дословного перевода (Direct Machine Translation). Суть данной категории заключается в минимальном преобразовании текста: в них операция перевода требует минимума операций с входными данными: исходный текст постепенно превращается в текст на выходном языке путем замены всех его элементов, найденных в словаре, на переводные эквиваленты.

Вторая подкатегория – трансферные системы (Transfer-based Machine Translation (TBMT). Основа данной подкатегории – синтаксическая или синтактико-семантическая структура (или несколько вариантов такой структуры). Перевод структур осуществляется не прямым способом, а через построение для каждого предложения синтаксической или синтактико-семантической структуры. Этапы анализа и синтеза в них независимы: анализ, как правило, многовариантный, ведется в категориях входного языка, синтез – и категориях выходного. Связь обоих этапов обеспечивается третьим компонентом – этапом межъязыковых операций (трансформаций), это собственно перевод, или трансфер.

Третья подкатегория – интерлингвистические системы (Interlingua Machine Translation). Как предполагалось, эта технология позволит восполнить недостаток семантики при переводе путем создания специального метаязыка, семантического языка-посредника, универсального для разных пар естественных языков.

Наиболее важным достижением данной подкатегории является модель «Смысл ↔ Текст», представляющая собой многоуровневую модель, которая позволяет перейти от текста к его смысловой структуре, записанной на некотором универсальном языке, после чего совершить обратный переход от записанной смысловой структуры к любому естественному языку.

Статистические модели перевода – Statistical Machine Translation (SMT) – возникли в первую очередь из-за невозможности описания языка только на основе правил. Это привело к созданию моделей, которые будут основаны на вероятностях и статистике, а не на грамматике.

Данная модель базируется на анализе большого количества данных и содержит два фундаментальных блока: модель перевода и модель языка. В качестве обучающих данных используется множество параллельных текстов, переведенных людьми как минимум на два языка – параллельные корпуса текстов.

Гибридная модель перевода – Hybrid Machine Translation – модель перевода, основанная на совмещение методов SMT и TBMT моделях. Методика применения гибридной модели заключается в использовании двуязычной базы часто встречающихся предложений – Translation Memory (TM). Данная система базируется на сравнении переводимых документов с данными, хранящимися в заранее созданной базе переводов. Такой подход позволяет не переводить один и тот же текст дважды, а обращаться к соответствующим сегментам в базе переводов, которые были выполнены ранее.

Нейронный перевод – Neural Machine Translation (NMT) – на данный момент является относительно новой и бурно развивающейся технологией. В основе системы нейронного машинного перевода лежат две основные идеи – рекуррентные нейронные сети и кодировка. Обычная нейронная сеть – это обобщенный алгоритм машинного обучения, который принимает список чисел и вычисляет результат.