ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Т. В. Бусел

МАШИННЫЙ ПЕРЕВОД: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Актуальность изучения и решения проблем, связанных с машинным переводом, и важность его практического применения в преодолении языкового барьера обусловлена значительным увеличением объема переводимых документов, нехваткой профессиональных переводчиков для многочисленных специализированных областей и отдельных языковых пар. В статье раскрывается суть работы современных систем машинного перевода. Рассматриваются три принципиально разных подхода к построению алгоритмов машинного перевода (основанный на правилах, статистический и гибридный), раскрываются их основные характеристики, преимущества и недостатки.

С середины XIX века, 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» [1, с. 269].

Проблема МП довольно сложна и не теряет своей актуальности и в XXI веке по целому ряду причин [2; 3; 4]. 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» [1, с 245].

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 вообще и МП в частности привлекает внимание теоретиков, в результате чего продолжают возникать все новые теории автоматизации перевода и формализации языковых данных и процессов [4].

При моделировании в автоматизированной системе *перевод* рассматривается как «многоуровневый процесс, где каждая процедура переводит компонент специального уровня» [5, с. 151]. 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» [5, с. 152]. 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 [4; 6].

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.

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.

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).

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 [3; 4; 6; 7]:

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* (рис. 1) 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 определѐнному количеству предшествующих слов в данном языке). Данная модель предоставляет возможности улучшить перевод, используя наиболее частотные словоупотребления на различных языках. Применение методов статистического перевода позволило успешно решить проблемы снятия смысловой многозначности, разрешения проблемы анафор (например, интерпретация местоимений), сегментации дискурса и др.

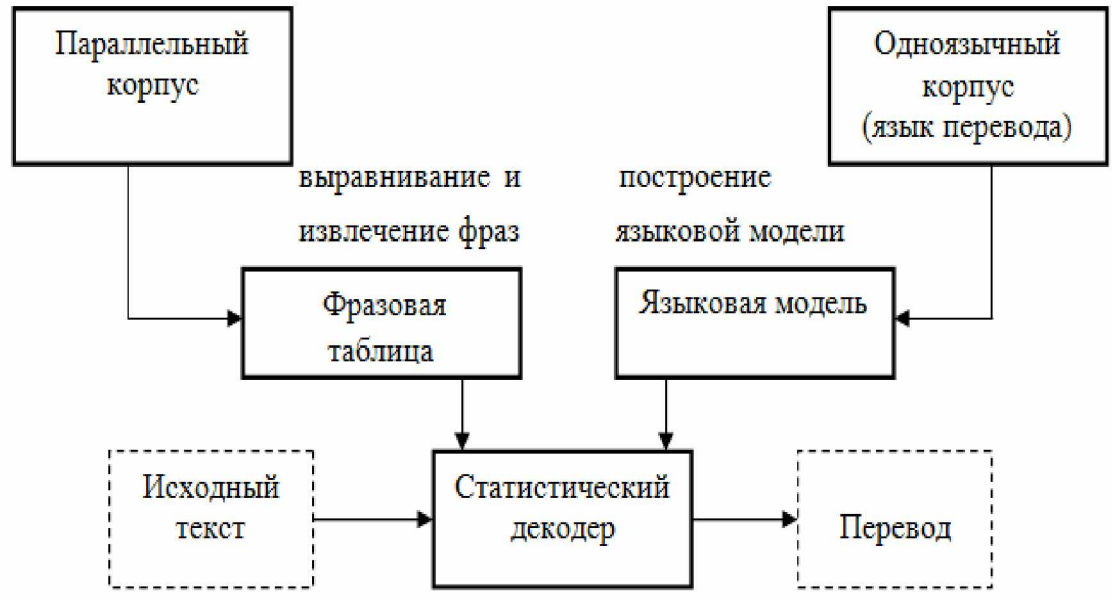

Рис. 1. Схема процесса перевода текста статистической системой

Системы на основе правил (рис. 2) производят анализ текста, который используется в процессе перевода. Перевод производится согласно встроенным словарям для данной языковой пары, а также грамматикам, охватывающим семантические, морфологические, синтаксические закономерности обоих языков. В соответствии со всеми этими данными исходный текст последовательно, предложение за предложением, преобразуется в текст на требуемом языке. Основной принцип работы таких систем – связь структур исходного и конечного текстов.

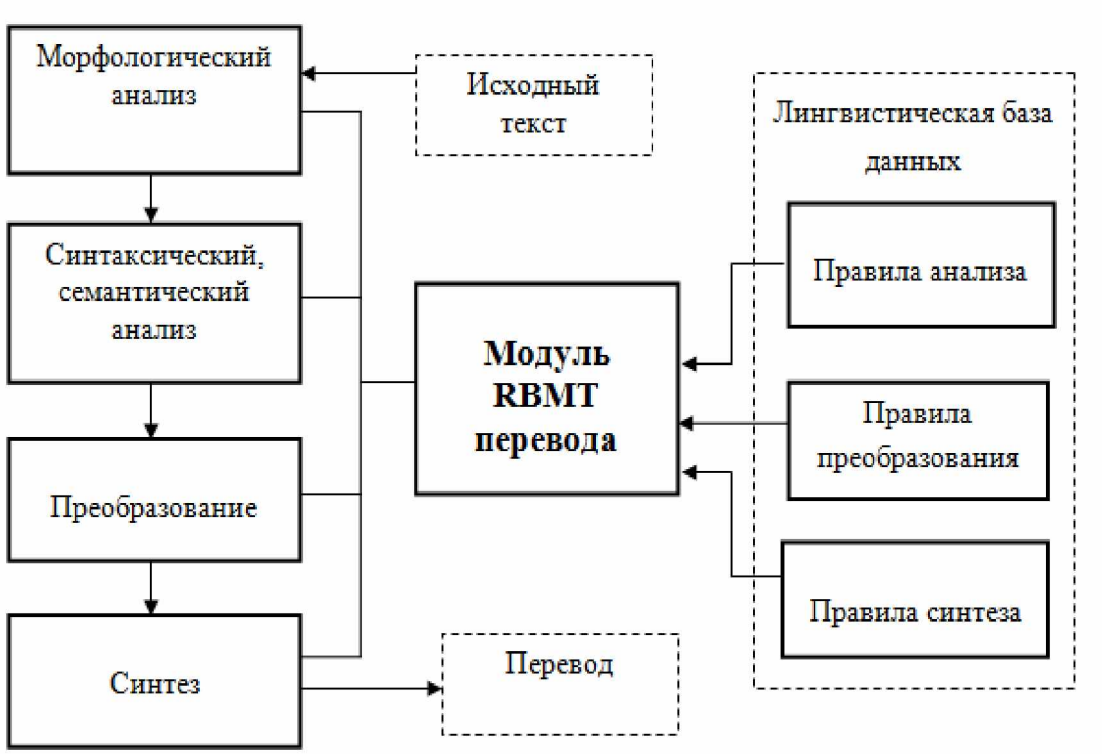

Рис. 2. Схема процесса перевода текста системой на основе правил

Среди систем МП, основанных на использовании правил, наиболее эффективными являются *трансфертные системы* [7], работающие 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» [7, с. 9]. На заключительном этапе синтеза формируется конечное предложение на языке перевода.

Основными преимуществами трансферных систем являются относительная гибкость лингвистического обеспечения, возможность добавления в систему новых правил без нарушения структуры всей системы; более высокое качество перевода, достигающееся за счет применения более разви-

тых формальных грамматик; более четкое регламентирование операций, выполняемых на каждом этапе перевода; ориентация на прагматическое описание естественных языков и, как следствие, относительно приемлемое качество перевода.

Гибридные системы машинного перевода (рис. 3) объединяют технологии машинного перевода на основе правил и на основе статистических моделей, выполняют лингвистический анализ входного предложения, порождение вариантов перевода, применение статистических технологий, оценку и выбор лучшего варианта перевода с использованием модели языка.

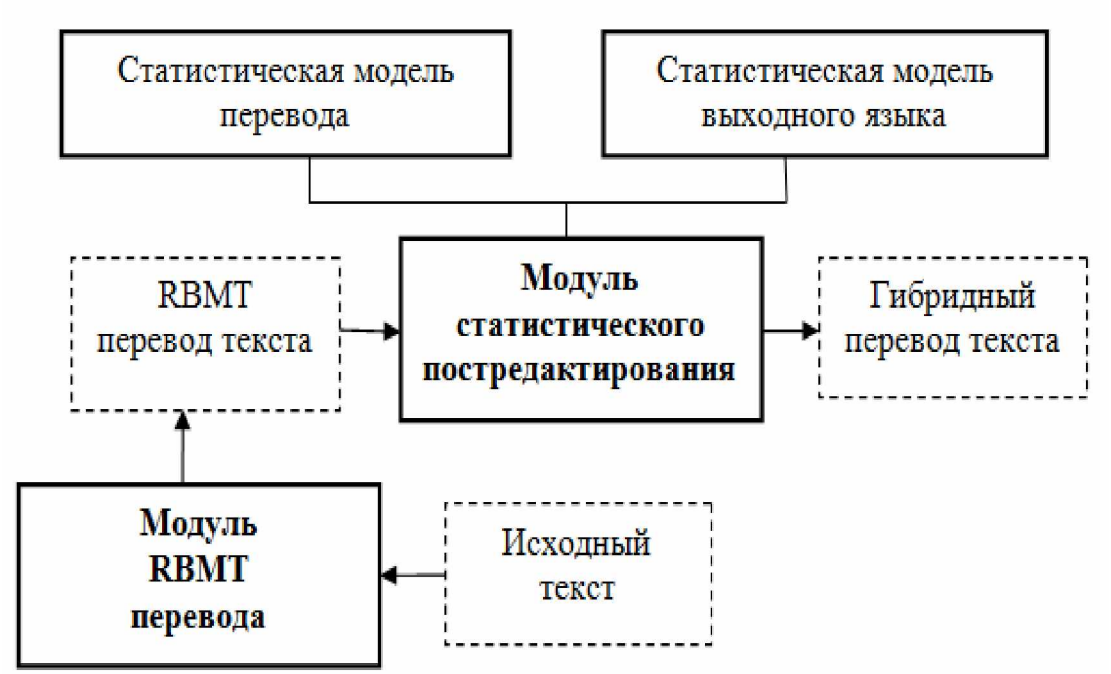

Рис. 3. Схема процесса перевода текста гибридной системой

Мы решили протестировать англо-русское направление перевода как наиболее популярное среди русскоязычных пользователей. В качестве примеров были взяты материалы американского издания “The Washington Post” и британского издания “The Gardian”.

Рассмотрим фрагмент перевода, выполненный системой «Google-Translate», которая основана на статистическом анализе: система подбирает эквивалент перевода в соответствии с частотой употреблений и в итоге подставляет вариант, имеющий наиболее высокий процент совпадений.

Как известно, для каждого языка характерны свои особенности лексико-семантической сочетаемости. Перевод, выполненный программой Google Translate, свидетельствует о том, что система не согласовывает слова друг с другом, употребляя практически все распознанные ей существительные в именительном падеже (за редким исключением) (табл. 1).

Таблица 1

Оригинал (фрагмент статьи “As Momentum Builds toward Tax Reform, Lobbyists Prepare for a Fight”)	Перевод, выполненный американской программой Google Translate	Перевод, выполненный переводчиком
While the standoff over sequester spending cuts and other budget battles have been grabbing headlines, momentum has quietly been building toward a once - in a -generation push to overhaul federal taxes, an effort that would likely affect nearly every family and business	*В то время как противостояние над поглощения расходов порезы и другие бюджета сражения захвата заголовки, импульс строит тихо направлении один раз в поколение push для ремонта федеральных налогов, что скорее всего скажется почти каждая семья и бизнес.*	*В то время, когда заголовки кричали о противостоянии по вопросу сокращения государственных расходов и других баталиях по поводу бюджета, постепенно усиливалось случающееся один раз в поколение давление на реформирование федеральных налогов, мера, которая, скорее всего, отразится на каждой семье и на бизнесе.*

Существительное *cuts* было переведено компьютерной системой как *порезы*, однако в статье речь идет о сокращении расходов, но для того, чтобы правильно передать данное словосочетание на русский язык, компьютерная программа должна поменять местами существительные, а этого она делать не умеет. Вспомогательные глаголы *have been* программа просто проигнорировала, отсюда и перевод существительного *захват* вместо глагола *захватывать*. Любой переводчик-человек понимает, что наречие *guietly* в данном контексте лучше переводить не *тихо*, а *спокойно*, *постепенно*, *не привлекая всеобщего внимания*. Лексему *push* система машинного перевода оставила без изменения, скорее всего, вследствие большого числа значений, предлагая человеку самому найти подходящее. Нелепо выглядят фразы *для ремонта федеральных налогов, скорее всего скажется почти каждая семья и бизнес*. Следует также отметить, что компьютерная система не расставляет знаки препинания. Согласно орфографии русского языка, словосочетание *скорее всего* должно выделяться запятыми.

Рассмотрим еще один пример работы автоматического переводчика PROMT, который действует по принципу «перевода по правилам», то есть работает по алгоритму, в соответствии с которым система анализирует текст на исходном языке и на основе проведенного анализа синтезирует перевод (табл. 2).

Таблица 2

Оригинал (фрагмент статьи “Venezuela Sets Date for Presidential Election”)	Перевод, выполненный российской программой PROMT	Перевод, выполненный переводчиком
Observers voiced mounting concern about the deep political divide gripping Venezuela, with half of it in a near frenzy of adulation and the other feeling targeted.	*Наблюдатели выразили обеспокоенность монтажа в глубокие политические разногласия, сцепление Венесуэлы, с половины его в ближайшем исступлении лести и другой для чувства целевых.*	*Наблюдатели выразили растущее беспокойство по поводу глубокого политического раскола, царящего в Венесуэле, одна половина которой охвачена безумным низкопоклонством, а другая считает, что её используют.*

Очевидно, что программа PROMT рассматривает выражение *mounting concern* как единое целое и при переводе даже поменяла слова местами. Не поняла она, однако, что это выражение представляет собой устойчивое словосочетание со значением ‘растущее беспокойство’, и ни о каком монтаже речи быть не может. Лексему *gripping* все-таки следует переводить как причастие в значении ‘царящий’, ‘приковывающий к себе внимание’, а не как технический термин *сцепление*. Причастие *targeted* образовано от полисемантичного глагола *target*, у которого есть значения ‘ставить или намечать цель’, ‘иметь целью’, ‘обстреливать цель’, но у всех этих значений есть помета – «военное дело». Общие (нейтральные) значения глагола – ‘намечать’, ‘планировать’, ‘делать кого-либо мишенью’, ‘выявлять’, ‘ориентировать’. При переводе следует отталкиваться именно от этих значений, что компьютерная система понять не в состоянии. Полисемантичным является и существительное *adulation*. Программа PROMT не умеет выбирать правильное значение, в данном случае *низкопоклонство,* а не *лесть*.

Полученные в результате анализа данные позволяют нам утверждать, что полностью автоматизированный качественный перевод на сегодняшний день невозможен. Результаты машинного перевода могут быть использованы лишь «для поверхностного ознакомления с содержанием при условии, что текст используется как сигнальная информация и не требует тщательного редактирования» [8, с. 19]. Улучшение качества современного машинного перевода представляет собой трудоёмкую 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.

ЛИТЕРАТУРА

1. *Марчук, Ю. Н.* Компьютерная лингвистика: учеб. пособие / Ю. Н. Марчук. – М. : Восток-запад, 2007. – 317 с.
2. *Hager, A.* The translation market in ten years' time – a forecast / A. Hager // TC WORLD magazine for International Information Management. – Stuttgart, 2008. – № 11. – S. 14–16.
3. *Мюге, У.* Три мифа о машинном переводе / У. Мюге // Профессиональный перевод и управление информацией. – 2009. – № 1(24). – С. 3–8.
4. *Марчук, Ю. Н.* Модели перевода : учеб. пособие / Ю. Н. Марчук. – М. : Академия, 2010. – 176 с.
5. *Беляева, Л. Н.* Теория и практика перевода: учеб. пособие / Л. Н. Беляева. – СПб. : ООО «Книжный Дом», 2007. – 212 с.
6. *Молчанов, А.* Статистические и гибридные методы перевода в технологиях компании PROMT / А. Молчанов // Control Engineering Россия. – 2013. – № 4(46). – С. 69–72.
7. *Новиков, В. А.* Трансфер в современных системах машинного перевода : автореф. дис. ... канд. филол. наук : 10.02.21 / В. А. Новиков. – М., 2001. – 25 с.
8. *Хроменков, П. Н.* Анализ и оценка эффективности современных систем машинного перевода : автореф. дис. ... канд. филол. наук : 10.02.21 / П. Н. Хроменков. – М., 2000. –21 с.

The article is devoted to modern machine translation systems. Today it is impossible to overestimate the significance and prospects of the machine-aided translation due to the rising demand for translations that is intensified by an existing lack of translators for numerous specialized fields and particular language combinations. The author reveals the essence of the work of modern machine translation systems of three types: rule-base, statistical and hybrid, and describes their main characteristics as well as advantages and disadvantages.

Поступила в редакцию 22.03.18