УДК 004.9 : 811.35

Т.А. Жданова,
канд. пед. наук, доцент,
доцент кафедры информатики
Тихоокеанского государственного университета
Г.А. Ефремова,
ассистент кафедры лингвистики и межкультурной коммуникации
Тихоокеанского государственного университета
(г. Хабаровск)

АВТОМАТИЗАЦИЯ ПЕРЕВОДА: ПРОШЛОЕ И НАСТОЯЩЕЕ

Авторы статьи рассматривают исторические аспекты машинного перевода и анализируют современные технологии.

Ключевые слова: *кодирование языков, машинный перевод, Джорджтаунский эксперимент, статистические системы.*

The authors consider historical aspects of machine automatic translation and analyze modern technologies.

Keywords: *languages' coding, machine automatic translation, Georgetown experiment, statistical systems.*

Проблема кодирования языков в целях обеспечения их доступности иноземцам встала перед представителями различных народов еще в древние времена. В XVII в. была предложена идея применения специальных машин для перевода слов с одного языка на другой.

Впервые к решению данной проблемы с научной точки зрения подошел Чарльз Бэббидж. В период с 1836 по 1848 гг. он спроектировал первую цифровую вычислительную машину, способную выполнять аналитические операции. Его идея заключалась в использовании памяти для хранения словарных баз [1]. К сожалению, проект по созданию прототипа системы Бэббиджа так и не был реализован.

В 1947 г. криптограф Уоррен Уивер изложил идею рассмотрения перевода в качестве новой области использования технологий декодирования. Уивер утверждал, что для перевода текста с одного иностранного языка на другой необходимо лишь представить, что исходный текст написан на родном языке, но закодирован с помощью специальных знаков. Тогда, в условиях послевоенного времени, сравнение технологий перевода с дешифрованием казалось вполне естественным. Идея Уивера вызвала большой резонанс. В 1949 г. публикуется его меморандум, теоретически обосновавший возможность реализации предложенной идеи машинного перево-

да (далее – МП). В те годы началось рождение МП как самостоятельного научного направления.

Первый успех в сфере автоматизации перевода связан с Джорджтаунским экспериментом в 1954 г. на машине IBM-701. Программно-языковое обеспечение данного эксперимента было довольно скромным: словарная база из 250 слов и 6 синтаксических правил давали возможность перевести всего 49 предварительно отобранных предложений. Однако это положило начало настоящему исследовательскому буму.

В 50-х гг. прошлого века первые успехи в этой области привели к созданию нескольких глобальных проектов, направленных на решение проблемы МП. Главным образом решались вопросы моделирования языка и языковых аспектов, языковой и мыслительной деятельности, изучения языковой формы. Однако в 60-х гг., после публикации доклада специального комитета по прикладной лингвистике Национальной академии наук США, в котором утверждалось, что системы автоматического перевода не смогут обеспечить приемлемое качество переводов в будущем, финансирование подобных исследований было прекращено.

Тема МП вновь приобрела актуальность лишь на заре 80-х годов. Были выработаны способы морфологического анализа для основных языков Европы, сформулированы основные требования к семантическим элементам таких систем и разработаны методы автоматического анализа синтаксических структур. За это время были созданы крупные международные проекты с солидными бюджетами, к числу которых относятся EU ROTRA (Европейское экономическое сообщество), METAL (США и Германия), ARIAM (Франция), KANT (США) и пр. Но, как и раньше, ни один из этих проектов не смог предложить уникальное решение, применимое в рамках массового использования.

Глобальные проекты были посвящены решению общих задач МП и ориентированы на разработку способов описания слов, входящих в терминологическую базу словаря и отдельно на создание алгоритмов перевода. Существенным недостатком таких систем было ограниченное количество грамматических алгоритмов, которые можно было применить для описания лишь определенной части предложений, но на их основе невозможно было правильно анализировать и переводить реальные тексты. Несмотря на это, проведенные работы позволили понять всю сложность стоявшей перед разработчиками задачи. В дальнейшем именно такие проекты легли в основу систем машинного перевода, которые сегодня предлагаются пользователям. К их числу относятся системы Power Translator (производство Globalink), TRANSEND (компания Intergraph), а также Language Assistant (MicroTac). Однако и «вторая волна» разработок и исследований в области МП тоже сошла на нет. Сложность задач, с которыми столкнулись создатели МП, была выше, чем уровень развития аппаратно-программного обеспечения.

Зато 90-е годы, в течение которых

индустрия информационных и коммуникационных технологий пережила бурный прогресс, стали эпохой возрождения машинного перевода: создание персональных компьютеров, внедрение Интернета и локальных сетей обусловили стремительный рост интереса к машинному переводу.

В настоящее время имеется достаточно широкий выбор пакетов программ, которые условно можно разделить на две основные группы: электронные словари (electronic dictionary) и системы машинного перевода (machine translation system). Системы машинного перевода (далее – СМП) текстов с одних естественных языков на другие моделируют работу человека-переводчика. В данный момент выделяют три типа СМП.

Полностью автоматические системы машинного перевода являются, скорее, несбыточной мечтой. Все системы машинного перевода (МТ-системы) работают при участии человека в той или иной мере. ТМ-системы иногда называют еще «памятью переводчика». Они являются просто удобным инструментом, нежели элементом автоматизации. Сейчас заменить человека-переводчика машиной, конечно, невозможно, но можно значительно облегчить труд переводчика и повысить его производительность.

Современные системы машинного перевода используют определенные методы, в результате чего их можно разделить на три большие группы: 1) СМП, основанные на правилах; 2) СМП, основанные на примерах; 3) статистические СМП.

СМП, основанные на правилах, – общий термин, который обозначает системы МП на основе лингвистической информации об исходном и переводном языках, в основном полученной из двуязычных словарей и грамматик, охватывающих основные семантические, морфологические, синтаксические закономерности каждого языка. Такой подход к машинному переводу еще называют классическим. На основе этих данных исходный текст последовательно, по предложениям, преобразуется в текст перевода. Перевод при этом получается не особенно хорошего качества, но на простых примерах работает.

СМП, основанные на правилах, делятся на три группы: 1) системы пословного перевода (далее – СПП); 2) трансфертные системы (далее – ТС); 3) интерлингвистические системы (далее – ИС).

СПП используются сейчас крайне редко из-за низкого качества перевода. Слова исходного текста преобразуются (как есть) в слова переводного текста. Он используется для перевода длинных списков слов (например, каталогов). Также он может быть использован для составления подстрочечника для ТМ-систем.

Как ТС, так и ИС имеют одну и ту же общую идею. Для перевода необходимо иметь посредника, которой несет в себе смысл переводимого выражения. В ИС посредник не зависит от пары языков, в то время как в ТС – зависит. ТС работают по очень простому принципу: к входному тексту применяются правила, которые ставят в соответствие структуры исходного и переводного языков. При использовании этой страте-

гии получается достаточно высокое качество переводов, с точностью примерно 90 % (хотя это очень зависит от языковой пары). Работа любой ТС перевода состоит, как минимум, из пяти частей: 1) морфологический анализ; 2) лексические категоризации; 3) лексический трансфер; 4) структурный трансфер; 5) морфологическая генерация.

ИС машинного перевода – один из классических подходов. Исходный текст трансформируется в абстрактное представление, которое не зависит от языка. В рамках такого подхода можно реализовать «пересказ текста»; перефразирование исходного текста в рамках одного языка; относительно простой перевод сильно отличающихся языков, таких как, например, русский и арабский. Однако до сих пор не существует реализации такого подхода, которая бы корректно работала хотя бы для двух языков. Многие эксперты высказывают сомнения в возможности такой реализации. Сама большая сложность для создания подобных систем заключается в проектировании межъязыкового представления. Оно должно быть одновременно абстрактным, не зависящим от конкретных языков, но в то же время отражать особенности любого существующего языка.

Перевод, основанный на примерах, – один из подходов МП, при котором используется двуязычный корпус текста. Этот корпус текста во время перевода используется как база знаний. Примерами двуязычных корпусов текстов можно назвать парламентские отчеты в Канаде, Гонконге и других странах. Тексты представляют собой протоколы дебатов в парламенте. Кроме того, хорошим примером являются официальные документы на нескольких языках, которые также считаются полезными для МП. Предполагается, что люди разлагают исходный текст на фразы, потом переводят эти фразы, а далее составляют текст из фраз. Причем перевод фраз обычно происходит по аналогии с предыдущими переводами. Для построения СМП, основанной на примерах, потребуется языковой корпус, составленный из пар предложений.

Например: английский и японский (латиница).

How much is that red umbrella? Anoakai kasawa ikura desu ka?

How much is that small camera? Ano chiisai kamera wa ikura desu ka?

Языковые пары, содержащие предложения на одном языке и соответствующие им предложения на втором, могут быть как вариантами написания двух предложений человеком – носителем двух языков, так и набором предложений и их переводов, выполненных человеком. Можно заметить, что предложения из примера отличаются только двумя словами. И это можно видеть для обоих языков. В данном случае для перевода нужно знать только три единицы информации: *1. How much is that X? – Ano X wa ikura desu ka?; 2. red umbrella – akai kasa 3. small camera – chiisai kamera.*

В данном случае мы оперируем более высокоуровневыми данными. Перевод, основанный на примерах, лучше

всего подходит для таких явлений, как фразовые глаголы. Смысл такого выражения невозможно получить из смыслов составляющих частей. Классические методы перевода в данном случае неприменимы. Статистический МП – это метод МП, при котором используется сравнение больших объёмов языковых пар. Чем больше в распоряжении имеется языковых пар и чем точнее они соответствуют друг другу, тем лучше результат статистического МП. Суть такого перевода заключается в поиске вероятного перевода предложения с использованием данных из двуязычных корпусов текстов. В результате при выполнении перевода компьютер не оперирует лингвистическими алгоритмами, а вычисляет вероятность применения, того или иного слова или выражения. Слово или последовательность слов, имеющие оптимальную вероятность, считаются наиболее соответствующими переводу исходного текста и подставляются компьютером в получаемый в результате текст.

Работа статистических систем происходит в двух режимах – обучения и эксплуатации. В режиме обучения просматриваются параллельные корпуса текста и вычисляются вероятности переводных соответствий. Строится модель языка перевода. В режиме эксплуатации для фразы из исходного текста ищется фраза переводного текста так, чтобы максимизировать произведение вероятностей. Самой простой статистической моделью перевода является модель дословного перевода. В этой модели, известной как Модель IBM № l, предполагается, что для перевода предложения с одного языка на другой достаточно перевести все слова, а расстановку их в правильном порядке обеспечит модель языка [2].

В данный момент статистические системы являются лидерами по соотношению цена/качество для всех СМП. Представителями данных систем на рынке являются IBM Model 1-5, Google, Moses, Rewrite и некоторые другие.

Применяя МП, следует помнить, что машинный перевод изначально создавался для оперативного перевода технической документации, и потому именно в этой сфере он проявляет себя лучше всего и тогда, когда нужно в кратчайшие сроки получить общее представление о каком-либо явлении или изобретении, и можно пренебречь некоторыми деталями, и красотой языка.

Список использованных источников

1 Апокин И. А. / И. А. Апокин, Л. Е. Майстров, И. С. Эдлин. М. : Наука, 1981.

2 Рахимбердиев Б. Н. Эволюция семантики экономической терминологии русского языка в XX веке : дис. ... канд. филолог. наук / Б. Н. Рахимбердиев. М., 2003. 188 с.

3 http://www.studmed.ru/docs/document23659?view=6

4 http://bibliofond.ru/view.aspx?id=513466#1