

Роберт Шимуля
Белосток

Проблема устранения многозначности при машинном переводе

Машинный перевод представляет собой одну из самых интересных теоретических и практических проблем лингвистики. Нужно отметить, что десятилетия развития научных исследований и практических разработок внесли свой существенный вклад в теорию вопроса и в приложение его на практике.

Термин „перевод” можно понимать двояко: как процесс перехода от текста на одном языке к этому же тексту на другом языке или его результат. Существует большое количество видов переводов и их классификаций (в зависимости от переводного материала, по форме презентации текста перевода и текста оригинала, по цели, с какой делается перевод и т. п.), но для нас самой существенной является классификация по степени механизации процесса перевода. По этой классификации можно выделить традиционный перевод, а также три способа перевода с использованием компьютера, которые в зависимости от степени вмешательства человека в процесс перевода можно разделить на: перевод текста человеком с использованием компьютера, перевод текста компьютером с использованием человека, перевод текста компьютером без помощи человека (машинный или автоматический перевод)¹. Перевод с использованием компьютера можно разделить также в зависимости от того, насколько далекие цели ставят перед собой разработчики. В таком

¹ И. И. Зубова, *Информационные технологии в лингвистике*, Минск 2001, с. 87.

случае средства компьютеризации перевода подразделяются на три основных типа: системы машинного перевода, системы человеко-машинного перевода (которые подразделяются на системы машинного перевода с участием человека и системы человеческого перевода с участием машины) и терминологические банки данных².

Перевод текста человеком с использованием компьютера имеет место в тех случаях, когда перевод обеспечивается человеком, но в некоторых ситуациях человек может прибегать к помощи компьютера: при поиске слов в локальных словарях или тезаурусах, при дистанционном обращении к терминологическим банкам данных, при выборке примеров употребления слов и словосочетаний, хранящихся в памяти компьютера, при выполнении различных операций, связанных с экранным редактированием текста (например, операций форматирования) и т. п.

О переводе текста компьютером с использованием человека говорится применительно к тем системам, в которых перевод осуществляется компьютером, но на определенных этапах человек подключается к этому процессу – может выступать предредактором, который осуществляет подготовительную обработку текста, интерредактором, к которому компьютер в процессе осуществления перевода может обращаться за справками (может просить его определить часть речи или значение обнаруженного омонима, выбрать для данного слова или словосочетания один из нескольких переводных эквивалентов, находящихся в словаре системы, и т. п.), или постредактором, который осуществляет полную редакторскую правку переведенного автоматической системой текста.

Терминологические банки данных значительно отличаются от описанных выше систем. Они включают обычно только техническую терминологию, но не общеупотребительные слова, которые пользователь уже знает. Основное достоинство этих систем заключается в том, что они содержат самые последние данные, так как техническая терминология постоянно меняется и печатные словари часто являются в значительной степени устарелыми. Терминологические банки данных могут также превосходить обычные словари по объему. Обращение к ним происходит нередко не в ходе самого процесса перевода, но еще до того, как человек начнет свою работу.

² Дж. Слокум, *Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития*, „Новое в зарубежной лингвистике”, Москва 1989, Вып. XXIV, с. 357–407.

Машинный (или автоматический) перевод определяется как

выполняемое компьютером действие по преобразованию текста на одном естественном языке в текст на другом естественном языке при сохранении эквивалентности содержания, а также результат такого действия³.

Такое понимание этого типа перевода не исключает ни предварительной подготовки текста, ни постредактирования (постредактирование используется также при традиционном переводе, выполняемом человеком). Главный отличительный признак машинного перевода – это факт, что весь процесс перевода целиком обеспечивается самой системой без какого-либо участия человека, с использованием только специальных программ, компьютерных словарей и наборов лингвистических правил. К. Буатэ разделяет еще машинный перевод на три основных типа: „информативный” (грубый пословный перевод, достаточный для поверхностного ознакомления с содержанием текста на незнакомом языке), „профессиональный” (которого качество сравнимо с качеством перевода, сделанного человеком) и „персональный” машинный перевод (в этом случае авторы текстов, подлежащих переводу, заранее избавляют их от неоднозначностей и работают в режиме диалога с компьютером, что позволяет значительно улучшить качество перевода)⁴.

Область машинного перевода имеет достаточно длительную историю. Первый эксперимент по машинному переводу был проведен в США в университете в Джорджтауне 7 января 1954 года. В России первый машинный англо-русский перевод был выполнен в 1955 году. В течение первых лет интерес к этой проблеме стимулировался надеждой на возможность скоростного высококачественного перевода текстов любого типа⁵. Уже во втором десятилетии начали появляться признаки разочарования, усиливавшиеся по мере того,

³ *Научно-технический перевод*, Москва 1987, с. 80.

⁴ K. Boitet, *Twelve Problems for Machine Translation*, [w:] International Conference on Current Issues in Computational Linguistics. Univ. Sains Malaysia, Penang, Malaysia, 1991. Proceedings. P. 45–57. Сит. за. за: Ю. Н. Марчук, *Модель „текст-текст” и переводные соответствия в теории машинного перевода*, „Проблемы компьютерной лингвистики”, Минск 1997, с. 25.

⁵ Стоит подчеркнуть, что это требование необходимо пересмотреть. Для систем машинного перевода важны только два элемента: способна ли система давать на выходе результаты, приемлемые по своему качеству для тех целей, для которых они предназначены и является ли вся работа в целом экономически рентабельной.

как выявлялись все новые лингвистические проблемы и все яснее осознавалось, что процесс перевода отнюдь не так легко поддается автоматизации, как это вначале думалось. Неудача машинного перевода в тот период была не случайной: большинство ученых имело доступ к вычислительным машинам, память которых была недостаточна для работы, а ввод и вывод текста был связан с трудоемким процессом перекодировки. Кроме того, отсутствовали кадры инженеров-программистов, а вместе с ними и понимание того, что машинный перевод является не только теоретической проблемой, но также инженерной задачей, для решения которой требуется техническая база и умелое сочетание теоретической мысли и практической организации лингвистических, математических и программистских исследований.

Актуальность автоматизации перевода продолжает оставаться высокой. Все время растут и расширяются возможности современной компьютерной информационной технологии, благодаря чему эффективный машинный перевод становится все более реальным. Интерес к этой идее вызван в большой степени возможностью практического применения, хотя высока и ее научная привлекательность. Следует подчеркнуть, что спрос на машинный перевод литературных произведений практически отсутствует, так как системы машинного перевода плохо справляются с такой задачей. Зато спрос на перевод разного рода технических текстов быстро увеличивается соответственно тому, как все больше естественных языков вступает в коммуникационную информационную сферу. При большом объеме требуемых переводов на тщательную разработку каждого документа традиционным способом не хватает ни времени, ни средств (квалифицированных переводчиков технических текстов найти трудно, они работают медленно, и стоимость их работы весьма высока). Благодаря использованию разного рода автоматизированных систем перевода (как систем машинного, так и систем человеко-машинного перевода) увеличение производительности человеческого труда стало возможным.

Кроме быстроты работы, возможность использования такого рода систем обуславливается также тем, что основной задачей при переводе технического текста является адекватное отражение предметно-понятийной стороны оригинала, и упор делается на точность в отличие от перевода художественной литературы, в котором первостепенное значение придается стилистическим требованиям, даже в ущерб требованию абсолютной точности в передаче содержания.

Именно поэтому автоматизированные системы в наибольшей степени пригодны для технического перевода – с точки зрения стилистики они неудовлетворительны, но с точки зрения терминологии являются очень хорошим орудием. Важна также возможность их использования при дальнейшем распространении информации, например, при экспорте технического оборудования. Если фирма заинтересована в продаже продукции за пределами своей страны, ей обычно приходится предоставлять покупателю документацию на том языке, который он предпочитает. Новый подъем интереса к автоматизированному переводу во многом вызван именно перспективой его применения для таких целей.

Другой областью, в которой качество автоматизированного перевода вполне приемлемо, является научная литература. Так, количество материалов на иностранных языках в сфере технических текстов, с которыми нужно ознакомиться ученым, превышает их возможности. Кроме того, ученым, которые ограничиваются только английским языком, который когда-то преобладал в научном общении, сейчас из-за распространения передовых технологий по всему миру все труднее держаться в курсе последних достижений в своей области знаний. В такой ситуации даже низкокачественные переводы литературы из области, в которой данный ученый является специалистом, служат хорошим средством быстрого сбора информации и позволяют ему понять содержание текстов достаточно хорошо, чтобы обойтись без тщательного традиционного перевода.

Несмотря на значительное число работ по машинному переводу, многие лингвистические вопросы в настоящее время все еще остаются актуальными. Одной из таких основных проблем является задача снятия неоднозначности лексических единиц в двуязычной ситуации⁶. Двуязычная ситуация характеризует особый план – план соотношения между лексическими единицами разных языков. Двуязычную неоднозначность можно определить как

одно-многозначное соотношение между означающим входного слова и означающими слов выходного языка, соотносимых с входным словом при переводе⁷.

⁶ У большинства систем затруднение вызывает перевод многозначных существительных, и в таких ситуациях программа часто выбирает первичные значения слов, которые наименее обусловлены контекстом.

⁷ Л. И. Трибис, *Устранение двуязычной неоднозначности в системе „человек-машина-человек“*, „Вопросы филологии“, Минск 1973, Вып. 3, с. 168–169.

В конкретном контексте речевого употребления неоднозначность, как правило, устраняется, т. е. приводится к однозначности. При этом контекст не выступает однородной средой, в нем различимы формально выраженные, диагностирующие признаки, однозначно определяющие выбор соответствующей выходной единицы. Суть процесса перевода неоднозначной входной единицы состоит в нахождении диагностирующих признаков, детерминирующих реализуемый в данном контексте переводный эквивалент. При машинном переводе задача заключается в создании фильтров для устранения двуязычной неоднозначности, т. е. выбора из множества означающих переводных эквивалентов входного слова перевода, употребляемого в данном случае.

Ученые по-разному подходят к решению этой проблемы. Одним из выходов может служить использование перевода текста компьютером с использованием человека вместо „чистого” машинного перевода. Суть такого решения заключается в использовании предредактора, готовящего текст таким образом, чтобы облегчить машине его анализ, или постредактора, вычеркивающего ненужные значения, выбранные компьютером. Следующий способ заключается в соотношении слов с определенными семантическими полями. Эту идею теоретически разработали В. Ингве и Л. Достерт, но ее практическая реализация оказалась невозможной. Чтобы избежать полисемии, предлагалось также подавать текст определенным способом, т. е. писать на всех языках имена существительные с заглавной буквы, в глаголах делать заглавной вторую букву, в прилагательных третью и т. д., с целью облегчить машине восприятие текста.

Одним из наиболее разработанных является контекстуальный подход, заключающийся в снятии многозначности после анализа окружения конкретного неоднозначного слова. Его практическая реализация сводится к анализу классов слов, окружающих лексическую единицу. Л. А. Славина и Л. И. Трибис на основании проведенного исследования, реализовавшего контекстуальный подход, называют три типа признаков, являющихся диагностирующими для каждого переводного эквивалента: лексические диагностирующие признаки (лексические группы, наполняющие модели сочетаемости анализируемых слов или отдельные сочетающиеся словоформы), морфологические диагностирующие признаки (суффиксы как носители определенной семантической информации о самом неоднозначном входном слове или сочетающихся словах), формальные диагностирующие признаки (написание прописными буквами, матема-

тические формулы, цифровая запись и т. д.)⁸. В проведенном опыте анализ списков типовых контекстов для каждого анализируемого слова обнаружил, что неоднозначность исследуемых входных слов устраняется обращением к контексту для учета перечисленных диагностирующих признаков в пределах именных и глагольных сегментов, синтаксически связанных с исследуемыми словами. Лексическое детерминирование переводных эквивалентов неоднозначных входных слов происходило при помощи списков диагностирующих словоформ. В большинстве случаев для исследуемых неоднозначных слов диагностирующий признак находился внутри нулевого (т. е. того, в котором находилось анализируемое неоднозначное слово) или контактирующих с ним сегментов, реже – в неконтактирующих сегментах. В связи с этим во время анализа поиск диагностирующих признаков мог организоваться только по нулевому, первому правому и первому левому сегментам почти во всех случаях. Работа составленных алгоритмов устранения двуязычной лексической неоднозначности осуществлялась как комбинации различного рода стандартных операторов – проверок на наличие диагностирующих признаков в автоматически извлекаемых из текста сегментах анализируемых предложений. Выбор нужного эквивалента из гнезда осуществлялся в результате анализа контекста (поиска определенных диагностирующих признаков в предусмотренных алгоритмом сегментах).

По-другому неоднозначность устраняется при так называемом тезаурусном методе. Семантический анализ текстов понимается здесь как переход от слов текста к индексам тематических групп, причем выбирается тот из индексов, который приписывается другим словам данного текста. Семантический синтез производится путем отыскания для формулы, полученной в результате анализа, различных возможных выражений в тематических группах тезауруса выходного языка. Неоднозначность при таком подходе состоит в том, что один и тот же входной термин попадает в разные узлы тезаурусной иерархии, в результате чего он получает не один, а несколько наборов тезаурусных кодов. Задача процедуры снятия тезаурусной многозначности состоит в том, чтобы путем анализа тезаурусных

⁸ Л. А. Славина, Л. И. Трибис, *Об алгоритмическом устранении двуязычной лексической неоднозначности*, [w:] *Частные вопросы автоматического анализа текстов*, Минск 1972, с. 295.

кодов текстового термина определить его точное положение в тезаурусе, а затем, учитывая окружение термина, дать правильный перевод. Такой контекстный семантический анализ должен привести к снятию лексической многозначности.

В последнее время возрастает интерес к разрешению проблемы полисемии слов путем использования параллельных корпусов текстов, которые можно определить как достаточно большие массивы оригинальных текстов и их переводов на другой язык, обладающих лингвостилистической и смысловой адекватностью по отношению к оригиналам. Параллельные тексты могут использоваться для решения многих лингвистических задач, например, в лексикографии и лексикологии (для составления словарей, определения значений многозначных слов и т.п.) или в грамматике (для определения частоты употребления грамматических морфем в текстах различного типа, выявления наиболее употребляемых типов словосочетаний и предложений и т.п.). В области автоматического перевода параллельные тексты могут применяться для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов терминологических и фразеологических словосочетаний и для других целей. Лингвисты основываются на положении, что для достижения качественного перевода невозможно обойтись без знаний, навыков переводчика-человека. Человек, переводящий текст с одного языка на другой, творчески анализирует материал. Его мозг отбирает и ассоциирует признаки воспринимаемой информации, находит среди них такие, которые определяют выбор уместного в данном контексте перевода многозначной единицы. Человек ориентируется как на знание признаков изучаемого материала, так и на интуицию, позволяющую ему пересматривать варианты перевода и останавливаться на наиболее приемлемом. В связи с этим для нахождения верных эквивалентов в языке перевода лексическим единицам в исходном языке необходимо обратиться к уже готовым переводам текстов различной тематики. Следует отметить, что важную роль при эффективном использовании параллельных текстов для реализации машинного перевода играет количество и объем самих текстов. Чем больше массив текста, тем больше он содержит переводных эквивалентов для единиц исходного языка.

В последние годы для устранения многозначности при машинном переводе стали использоваться также таггированные корпуса текстов. Все слова такого корпуса получают некоторые буквенные или цифровые индексы, которые обозначают их грамматические, ле-

ксические, семантические или структурные признаки. Таггированные тексты применяются, например, для установления принадлежности слова предложения к тому или иному грамматическому классу, определения частоты структуры предложений на уровне частей речи и т. д.

Машинный перевод является одной из самых интересных теоретических и практических задач современной лингвистики. Нет сомнений в том, что эта проблема будет исследоваться также в будущем. Несмотря на значительное число работ в данной области, многие вопросы все еще остаются актуальными. Одной из таких основных проблем является задача устранения неоднозначности лексических единиц в двуязычной ситуации. Принципы построения алгоритмов, служащих такой цели, обсуждались многократно, но эта проблема все еще нуждается в дальнейшем исследовании.

S U M M A R Y

The object of the analysis is the problem of multinomiality in automatic translation. The author presents a few methods of avoiding choosing not proper equivalents for multinomial words, concentrating on the contextual method and the method based on using parallel texts.