ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ
1958 МАТЕМАТИКА № 5 (6)

О. С. Кулагина

МАШИННЫЙ ПЕРЕВОД С ФРАНЦУЗСКОГО ЯЗЫКА

Работа по машинному переводу с французского языка на русский была начата в МИАН в конце 1954 года. Первый этап работы состоял в составлении алгоритма перевода и специального словаря. Алгоритм перевода составлялся совместно И. А. Мельчуком (ин-т Языкознания) и автором [1]. В составлении словаря принимали участие как математики, так и студенты-филологи (МГУ).

Начиная работу по машинному переводу, мы поставили себе задачу получить алгоритм для перевода текстов из области математики. Такое ограничение на тексты облегчало задачу перевода, так как в текстах указанного характера сравнительно ограниченный набор слов, конструкции фраз проще и не слишком разнообразны, кроме того, отсутствуют игра слов, прямая речь и т. п., которые особенно трудно перевести. Наша цель состояла в получении переводов, которые верно передавали бы смысл переводимого текста правильно сконструированными фразами.

Для машинного перевода был составлен небольшой словарь наиболее употребительных в математических текстах слов. Для этого был проанализирован словарный состав различных французских математических текстов. Было просмотрено 20500 слов, среди которых оказалось 2300 различных, из них отобрали слова, встретившиеся не менее четырех раз. К полученному добавили некоторые служебные слова и термины, необходимость которых была очевидна и которые не учитывались при просмотре текстов. Получили словарь из 1200 слов. Подсчитывались также так называемые обороты — словосочетания, которые не допускают дословного перевода (например, à peine — „едва", en effet—„действительно"). Они составили специальный словарь из 250 оборотов.

В словаре хранятся не слова целиком, а только их основы. Французские и русские основы в словаре снабжены рядом указаний, составляющих так называемую словарную информацию слов. Так для французского слова даются следующие указания: 1) указание части речи, 2) указание о вхождении слова в обороты, 3) указание на наличие омонимии или нескольких переводов, 4) указание, характеризующее индивидуальные особенности слов, если они имеются, 5) грамматические характеристики (например, род существительного, группа, переходность глагола и др.), 6) технические характеристики, нужные для упрощения работы программ перевода (например, число свободных разрядов в последней ячейке, занятой основой слова и др.).

[1] См. О. С. Кулагина и И. А. Мельчук. Машинный перевод с французского языка на русский. Вопросы языкознания № 5, 1956.

Для русского слова даются: 1) указание части речи, 2) указание о выборе основы (при наличии нескольких основ), 3) указание о типе изменения слова (склонения, спряжения), 4) технические характеристики.

Алгоритм перевода составлялся эмпирическим путем. Действия человека-переводчика при переводе некоторой фразы анализировались и записывались в виде правил. Эти правила применялись для перевода других фраз и в них вносились нужные поправки и дополнения до тех пор, пока не был получен устойчивый набор правил, позволяющий переводить фразы достаточно разнообразных грамматических конструкций. Разработанный алгоритм перевода рассчитан на перевод отдельными фразами; связи слов, лежащих в разных фразах, не анализируются.

Работа алгоритма происходит в следующем порядке: сначала работают правила поиска слов в словаре, затем правила обработки оборотов. После этого начинается различение омонимов. Под различением омонимов мы понимаем как разбор случаев, когда одно и то же слово может выступать в роли нескольких частей речи (омонимия 1 типа), так и выбор одного из нескольких переводов для слова, которое всегда является одной и той же частью речи, но имеет несколько значений (омонимия 2 типа или многозначность). Омонимы первого типа обязательно разбираются так, что часть речи устанавливается однозначно, так как оставление двусмысленности в этом случае затруднило бы анализ всей фразы, в то же время выбор одной из возможных частей речи может быть сделан на основе формального анализа конструкции фразы. Что касается разбора многозначности, то он производится не всегда, поскольку иногда не удаётся установить значение слова по формальным признакам, так как оно определяется только смыслом фразы. В таких случаях машина выдает несколько вариантов перевода.

После различения омонимов начинается анализ фразы по частям речи в следующем порядке: сначала анализируются все глаголы, затем предлоги, существительные, местоимения, причастия, не вошедшие в сложные формы глагола, и прилагательные.

Анализирующие правила устанавливают сведения о форме переводящего русского слова путем проверки ряда условий, касающихся либо формы французского слова, либо его положения среди других слов фразы. Например, может проверяться наличие у данного слова определенного окончания или наличие справа и слева от него определенных частей речи, индивидуальных слов.

По полученным при анализе данным синтезирующие правила конструируют русские слова и составляют из них переводящую фразу.

Составление алгоритма было в основном закончено в феврале 1956 г, после чего было начато составление программ и кодирование словаря. В кодировке участвовало 3 человека, в составлении программ — 8 человек.

Алгоритм перевода осуществлен при помощи 17 программ. Первая программа — программа поиска слов в словаре, она отыскивает для каждого слова переводимой фразы максимальную по длине основу, целиком умещающуюся в слове. Информация, стоящая в словаре при найденной основе, выбирается из словаря и запоминается в определенном участке памяти машины. Окончания — части слов, остающиеся после вычитания из слов основ, также запоминаются в определенном участке памяти. Все последующие программы уже не обращаются к словам фразы, а имеют дело только с извлеченной из словаря информацией и окончаниями.

Вторая программа обрабатывает обороты. Найдя во фразе обо-

рот, она заменяет найденные ранее словарные информации слов, составляющих оборот, информацией к обороту, взятой из словаря оборотов.

Третья программа — программа построения шкал. Дело в том, что на разных этапах работы выгодны различные способы хранения информации. В словаре информацию следует записывать так, чтобы она занимала мало места, а в процессе работы выгоднее такой способ хранения информации, который позволял бы легко с ней оперировать. С этой точки зрения оказывается удобным способ шкал, применявшийся и ранее в вычислительных задачах. Шкала для некоторого слова или группы слов — это ячейка, где единицы стоят в разрядах, номера которых совпадают с порядковыми номерами во фразе слов, которым отведена данная шкала. Третья программа строит шкалы для частей речи и для некоторых индивидуальных слов.

Программы с четвертой по шестую разбирают омонимы, затем семь программ анализируют французскую фразу и четыре программы синтезируют русскую фразу.

Общий объем программ составляет примерно 7000 приказов, кроме того, около 1800 ячеек занимают различные таблицы и константы. Словарь записан на магнитной ленте, он разбит на 26 зон и занимает около 12000 ячеек [1]. На перевод фразы в 8—10 слов машина затрачивает 45000—50000 тактов, из которых около 20000 приходится на долю программы поиска слов в словаре. На перевод фразы машина тратит 1,5—2 минуты.

Первый перевод фразы был получен в июне 1956 г. Правда, к этому времени не все программы были закончены, так не были доделаны программы различения омонимов. Осенью 1956 г. было кончено программирование и начато получение опытных переводов на машине. Работа по получению переводов и улучшению программ ведется Г. В. Чековой.

Фразы, переводимые на машине, были взяты из книг различных французских математиков: Borel, Picard, Bourbaki и др.

Приведу несколько примеров фраз, переведенных машиной [2].

1. Ensembles et éléments sont désignés dans les raisonnements par les symboles, qui sont en général les lettres ou les combinaisons de lettres.

Множества и элементы обозначаются в рассуждениях через символы, которые, вообще говоря, буквы или сочетания (комбинации) букв.

2. Les solutions précédentes tendent vers zéro quand t augmente indéfiniment.

Предыдущие решения стремятся к нулю, когда t возрастает неограниченно.

3. Les relations que nous avons trouvées entre les racines et les coefficients d'une équation conduisent assez naturellement á l'étude des formes symétriques.

Соотношения, которые мы нашли между корнями и коэффициентами уравнения, приводят достаточно естественно к изучению симметрических форм.

4 Le théoréme qui vient d'être établi subsiste dans ces nouvelles conditions.

Теорема, которая только что была установлена, существует в этих новых условиях.

5. Nous supposons que le cércle ait l'origine pour centre et l'unité

---

[1] Все приводимые здесь данные касаются первоначального варианта программ и словаря, в настоящее время они переделываются.

[2] Если слово имеет два перевода, то второй берётся в скобки.

pour rayon et de plus que le centre du cercle correspond au point Z_0 de l'aire A.

Мы предполагаем, что (чтобы) круг (окружность) имеет начало в качестве центра и единицу в качестве радиуса и сверх того, что (чтобы) центр круга (окружности) соответствует точке z_0 площади (области) A.

Рассмотрим теперь примеры фраз, первоначальный перевод которых нуждался в исправлении.

6. On peut conserver seulement deux membres de série (1).

Первоначальный перевод: „Мы **могем** сохранить только два члена ряда (1)".

Для перевода глагола „pouvoir" имеется как основа „мог", так и основа „мож", из-за ошибки в программе основа была выбрана неправильно. После исправления программы получили: „Мы можем сохранить только два члена ряда".

7. La considération d'une telle expression ne peut présenter aucun intérêt particulier.

Сначала был получен перевод: „Рассмотрение такого выражения не может настоящим никакой частный (особый) интерес".

В словаре есть основа „présent", которая может быть либо основой глагола (présenter — представлять), либо основой прилагательного (présent — настоящий). Этот случай не был учтён в правилах различения омонимов. После его добавления получили: „Рассмотрение такого выражения не может представлять никакой частный (особый) интерес".

По правилам после переходного глагола без отрицания существительное ставится в винительном падеже, а после отрицательного глагола — в родительном падеже. Но в данном случае, хотя ближайший к существительному глагол стоит в утвердительной форме, ему предшествует глагол с отрицанием, поэтому существительное должно также стоять в родительном падеже. После внесения этого правила получили перевод: „Рассмотрение такого выражения не может представлять никакого частного (особого) интереса".

Приведу еще примеры фраз, смысл которых передан верно, но перевод не вполне удовлетворителен по конструкции.

8. Cette transformation pourrait s'effectuer par les calculs relativement simples, en appliquant la remarque suivante.

Это преобразование сможет осуществиться относительно простыми вычислениями (исчислениями), применяя следующее замечание.

9. Nous signalerons seulement que lorsque on se place au point de vue de l'étude de ces correspondances, les fonctions d'ordre nul se divisent en deux grandes classes.

Мы отметим только, что (чтобы) когда мы помещаемся в точке зрения изучения этих соответствий функции равного нулю порядка делятся на 2 больших класса.

Дальнейшая работа в области перевода ведется в следующих трёх направлениях.

1. На основе полученных переводов исправляется и дополняется алгоритм и усовершенствуются программы перевода. Кроме того, переделывается словарь, так как оказалось целесообразным несколько изменить расположение данных в нем с тем, чтобы уменьшить его объём.

2. Разрабатывается формальная грамматическая система, которая позволила бы точно определять грамматические понятия и соотношения между ними [1].

---

[1] См. Бюллетень объединения по проблемам машинного перевода № 3. Стеклография, 1, МГПИИЯ.

Печатается в сб. „Проблемы кибернетики". Вып. 1.

3. Ведется разработка методов автоматизации работы по составлению программ перевода. Изучение имеющихся программ анализирующей части алгоритма перевода показало, что эти программы, несмотря на их кажущееся разнообразие, можно расчленить на сравнительно небольшое число элементарных актов переработки информации или операторов. Таких операторов в настоящее время выделено 17. По своим функциям эти операторы делятся на три группы. Первую группу составляют операторы, проверяющие некоторые условия. Во вторую группу входят операторы, производящие те или иные действия на основе данных, полученных при проверке. Это так называемые результирующие операторы. В третью группу входят операторы, не вошедшие в первые две группы или так называемые нейтральные. Приведем список операторов.

I. Операторы проверки

1. Проверка наличия у слова заданного окончания.
2. Поиск окончания в таблице.
3. Проверка наличия у слова определенной информации.
4. Поиск по информации слова с заданным признаком.
5. Проверка того, отмечено ли слово в заданной шкале.
6. Поиск по шкалам слова, отмеченного в определенной шкале.

II. Результирующие операторы

1. Запись информации.
2. Пометка слова в определенной шкале.
3. Стирание информации.
4. Вставка слова.
5. Передвижение информации.

III. Нейтральные операторы

1. Начальный.
2. Конечный.
3. Ветвление.
4. Формирование шкал.
5. Модификация.
6. Нестандартный.

Поясним, что нестандартный оператор — это условное название того куска алгоритма или программы, который не может быть записан в терминах остальных операторов.

Для того, чтобы некоторый оператор работал в нужном месте программы с нужным словом, ему должны быть заданы конкретные значения некоторых параметров. Например, для оператора поиска по информации надо указать: направление поиска, предмет поиска, от какого слова начинается поиск, какие слова можно пропускать при поиске, к какому оператору переходить, если нашли искомое, и к какому переходить, если не нашли искомое. Аналогично для каждого из остальных операторов должен быть задан определенный набор данных.

Работа по составлению программ перевода с помощью этих операторов будет идти следующим образом. Алгоритм перевода записывается в виде последовательности „простых“ правил, эти „простые“ правила являются либо проверками некоторых условий, либо действиями, совершающимися на основе результатов проверки.

К числу простых условий относятся проверка наличия у слова заданного окончания; проверка наличия окончания слова в некоторой таблице окончаний; проверка наличия у слова определен-

ного признака; поиск по фразе слова с определенным признаком в определенном направлении с определенными условиями пропуска слов при поиске (иначе говоря, проверка наличия или отсутствия такого слова в определенном куске фразы).

К простым действиям относятся запоминание каких-либо данных о слове; выбор одной из словарных информаций, если у слова их несколько; вставка слова; перестановка слов.

Легко заметить, что формулировки простых правил совпадают с формулировками некоторых операторов.

В записи алгоритма при каждом из простых правил должно быть указано, к выполнению какого правила следует переходить после выполнения данного. Если правило принадлежит к числу простых условий, то надо указать номер правила, к которому следует переходить, если условие выполнено, и номер правила, к которому следует переходить, если условие не выполнено.

Запись алгоритма в виде последовательности простых правил еще никак не связана с последующей реализацией алгоритма на той или иной машине, не требует никаких знаний о работе машин и может быть осуществлена составителем правил непосредственно при разработке алгоритма без какой-либо дополнительной затраты труда.

Полученная запись алгоритма превращается затем в операторную схему программы. Для этого каждое простое правило заменяется одним или несколькими операторами, его реализующими. Например, поиск слова по фразе может осуществляться оператором № 4 1-ой группы, либо оператором № 1 3-й группы, либо совокупностью операторов № 4 3-ей группы и № 6 1-ой группы. На этом этапе вводятся операторы, связанные с реализацией перевода в машине и не имеющие аналогов среди простых правил (такие как операторы № 5 2-ой группы и № 2 3-ей группы). Замена простого правила тем или иным оператором при наличии нескольких возможностей определяется набором шкал, размещением материала в памяти машины и до некоторой степени произвольна.

Операторы полученной схемы нумеруются. Для каждого оператора пишется его порядковый номер в схеме, затем его тип (номер в приведенном выше списке операторов), затем порядковые номера операторов, к которым следует переходить после работы данного, затем в определенной последовательности конкретные значения всех параметров данного оператора. Эта совокупность данных образует строку информации операторов.

Закодированная последовательность строк информации операторов вводится в машину и расшифровывается специальной компилирующей программой. Компилирующая программа имеет программные заготовки всех операторов, она выбирает из информации операторов значения параметров, вносит их в заготовки и строит из полученных кусков программу.

Приказы переадресации, передачи управления и другие, в которых фигурируют адреса команд строящейся программы, написаны в заготовках в некоторых условных адресах, компилирующая программа заменяет их на истинные адреса в зависимости от заданного ей начала строящейся программы.

Автоматизация процесса составления программ позволит сильно упростить и ускорить работу по осуществлению машинного перевода с одних языков на другие.

Математический институт
им. В. А. Стеклова
АН СССР