ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 5 1957

МАШИННЫЙ ПЕРЕВОД И ПРОБЛЕМА ЯЗЫКА-ПОСРЕДНИКА

Основным звеном машинного перевода является иероглифическое преобразование. Под иероглифическим преобразованием понимается выполняемое по заранее заданному алгоритму преобразование цифрового кода сообщения на одном языке в цифровой код того же сообщения на другом языке. Общая схема машинного перевода в принципе состоит из трех фаз: анализа, преобразования и синтеза. Анализ представляет собой кодирование сообщения, данного на входном языке (англ. «input language»), преобразование состоит в замене одного кода другим, синтез представляет собой раскодирование преобразованного сообщения в текст на выходном языке (англ. «output language»).

Практически в осуществленных до настоящего времени алгоритмах бинарного (т. е. с одного языка на один язык) перевода анализ и преобразование объединены. В этом случае, например, предложение на языке хиндустани *видъяртхӣ пустак пaрхтā хэй* получит следующую обработку при переводе на русский язык: *видъяртхӣ — студент*, *пустак — книга* (при отсутствии послелога переводится либо именительным, либо винительным падежом), *пaрх- — читать* (управляет винительным или дательным падежом), *-тā* — суффикс причастия наст. времени муж. рода ед. числа (переводится на русский язык причастием настоящего времени, если нет вспомогательного глагола *хōнā*; переводится настоящим временем знаменательного глагола, если есть форма вспомогательного), *хэй — есть* (3-е лицо ед. числа наст. времени вспомогательного глагола *быть*; не переводится после причастий настоящего времени). Составление русского предложения *студент читает книгу* согласно полученным данным осуществляется уже путем синтеза. Это же предложение при переводе на английский язык в слитном алгоритме обрабатывается иначе: *видъяртхӣ — student* (перед *student* надо добавить артикль *a*, если ранее это слово не упоминалось, артикль *the*, если слово ранее упоминалось), *пустак — book* (добавить артикль *a* или *the* на тех же условиях), *пaрх—read-*, *-тā—-ing*, *хэй — is*. Здесь ввиду отсутствия в английском языке падежных форм и параллелизма глагольного образования не нужны дополнительные указания, необходимые в предыдущем случае, но требуются иного рода дополнительные сведения, касающиеся введения артиклей. Для перевода на немецкий язык бинарная аналитико-преобразовательная часть примет третий вид, на французский — четвертый и т. д. Возможен, однако, и другой путь: закодировать в анализе хиндустани присущие ему грамматические категории для каждого слова, не равняясь ни на какой выходной язык, а затем включить эту одинаковую для всех случаев развертку в различные алгоритмы иероглифических преобразований. О третьем пути мы будем говорить далее.

При кодировании входного текста получаются цифровые знаки понятий, содержащихся в лексических единицах (семантические иероглифы), цифровые знаки грамматических морфем и знаки служебных слов (формальные иероглифы), цифровые знаки порядка слов и фонемно не выраженных синтагматических отношений между словами (тектонические иероглифы). При раскодировании соответствующие иероглифы определяют выбор слов для выходного текста, их грамматическое оформление и способы их сочетания.

И анализ, и синтез для каждого языка являются постоянными величинами, определяемыми исключительно нормами данного языка и принципами кодирования. В отличие от них иероглифическое преобразование — переменная величина, являющаяся функцией как от входной, так и от выходной иероглифики. Именно за счет переменности преобразования слитные алгоритмы перевода приведенного выше предложения с хиндустани на русский и с хиндустани на английский оказались столь различными. Что касается входной иероглифики, то она в обоих случаях была одною и той же.

В принципе возможны три типа несовпадения входной и выходной иероглифики: 1) при переводе необходимо убрать избыточный иероглиф (*die Sprache — язык*: избыточным является код артикля); 2) при переводе необходимо ввести дополнительный иероглиф (японск. *ани-но хон — книга брата*: дополнительными являются коды грамматических родов, отсутствующих в японском языке); 3) при переводе необходимо изменить тип иероглифа (*to catch cold — простудиться*: код одного из слов заменяется кодом глагольного вида, т. е. семантический иероглиф — формальным). Разумеет-

ся, эти несовпадения редко выступают в чистом, т. е. изолированном виде; чаще всего они комбинируются, что, собственно, и является основной причиной большой сложности алгоритмов перевода.

Условимся **неконгруэнтностью** называть любого рода несовпадение или неполное совпадение между входным и выходным комплексом иероглифов. Примеры неконгруэнтности формальных и тектонических иероглифов были даны выше. Еще бо́льшие трудности для машинного перевода составляет неконгруэнтность семантических иероглифов: англ. *board* — русск. *доска, стойка, контора, стол* (переносно, ср. *столоваться*) и, наоборот, русск. *стол* — англ. *table* — в прямом смысле, *board* — в переносном. Данные выполненных экспериментов машинного перевода не позволяют считать проблему преодоления этой неконгруэнтности окончательно решенной; в сущности она и не поставлена надлежащим образом, потому что определить конгруэнтный вариант значения полисемантического слова по контексту (как это делается в опубликованных алгоритмах) в большинстве случаев оказывается весьма сложной операцией для машины, не имеющей возможности обращаться к смыслу текста. Осуществление подобной операции потребовало бы наличия в «машинной памяти» обширной характеристики полисемантичного слова, которая должна содержать многочисленные указания на свободные словосочетания, позволяющие машине отбирать конгруэнтный вариант.

Мне представляется, что это — не лучший путь. Гораздо эффективнее была бы система семантических ключей. Эта идея основана на соображениях теории вероятности. Так, есть все основания ожидать, что в русской сельскохозяйственной статье слово *лук* будет означать огородную культуру, а не род оружия и что в астрономической статье слово *возмущение* значит изменение орбиты небесного тела под действием притяжения другого небесного тела, а не душевное состояние человека. Очевидно, что вторичные значения здесь не абсолютно исключены, но так же очевидно и то, что их встречаемость в подобных условиях близка к нулю, в силу чего процент ошибок, обусловленных пренебрежением вторичных значений, окажется в общем не выше обычного процента опечаток. Таким образом, для тех вариантов полисемантического слова, которые относятся к определенной лексической сфере, достаточно будет краткого условного ключа (типа словарных помет, употребляемых в лексикографической практике). Вместе с текстом, который следует перевести, машина получит его семантический ключ, который позволит машине сразу же отобрать внутри гнезд значений встречаемых в тексте слов именно те значения, которые являются конгруэнтными для данной специальной отрасли.

Следующим шагом, как мне кажется, должно быть отраслевое подразделение всего словаря в целом, т. е. выделение математической, химической, зоологической, музыкальной и др. терминологии в особые разделы со своим семантическим ключом каждый. Вводя в машину, скажем, математический текст, мы сопровождаем его ключом, который обязывает машину начать с перевода всей математической лексики, находящейся в статье, а потом уже обращаться к общему словарю.

Сам этот общий словарь — равно как и специальные словари — тоже может быть организован более экономично, чем это делалось до сих пор. Я имею в виду ступенчатую организацию, при которой словарь подразделяется на ступени: 1) наиболее употребительные слова, 2) слова средней употребительности, 3) редкие слова (такое подразделение словаря должно быть основано всецело на статистике). Машина вначале проверяет все неспециальные слова по первой ступени и, естественно, находит в ней большую часть таких слов, затем проверяет остальные по второй ступени и только с последними оставшимися словами обращается к третьей ступени. Легко показать, опираясь на формулы теории вероятности, что такая ступенчатая организация словаря в несколько раз выгоднее, чем осуществляемое в имеющихся алгоритмах сплошное пробегание единого алфавита слов, где перемешаны слова самой разной частотности.

Дальнейшее сокращение и упрощение словаря может быть достигнуто путем хотя бы частичной формализации словообразования. Сюда может быть включено типовое словосложение (*лесовоз, нефтевоз*; *машиностроение, станкостроение, компрессоростроение, вертолетостроение* и т. п.), при котором достаточно ввести в память машины семантические иероглифы типовых словоэлементов и тектонический иероглиф их положения в переводящем слове или словосочетании. Аналогичной трактовке может быть подвергнута типовая суффиксация, особенно развитая, например, в органической химии, где за каждым суффиксом закрепляется строго специализированное значение (ср. *эт-ан, эт-ил, эт-ил-ен, бут-ан, бут-ил, бут-ил-ен, проп-ан, проп-ил, проп-ил-ен* и т. д.). При этом десятки тысяч[1] таких терминов, как *дихлордифенилтрихлорэтан* (инсектицид, в просторечии именуемый дустом), сводятся к нескольким сотням иероглифов и, тем не менее, благодаря международности химической терминологии, конгруэнтность близка к 100%.

[1] Число соединений, известных органической химии, давно превысило 200 тысяч.

Введем понятие меры конгруэнтности, определив ее как частное от деления умноженного на сто количества конгруэнтных иероглифов на квадрат общего числа иероглифов, участвующих в преобразовании, т. е.:

$$K = \frac{100\,p}{(p+r)^2},$$

где p — число конгруэнтных иероглифов, r — число неконгруэнтных иероглифов, K — мера конгруэнтности [1].

Обозначим буквой σ_1 семантический иероглиф понятия «солнце», σ_2 — понятия «желтый», σ_3 — понятия «звезда»; φ_1 — формальный иероглиф глагола-связки, φ_2 — иероглиф мужского рода, φ_3 — женского рода, φ_4 — среднего рода, φ_5 — определенного артикля, φ_6 — неопределенного артикля, φ_7 — определительной частицы, φ_8 — 3-го лица ед. числа наст. времени глагола; τ_1 — тектонический иероглиф места сказуемого после подлежащего, τ_2 — иероглиф места определения перед определяемым, τ_3 — иероглиф места определения после определяемого, τ_4 — места артикля перед существительным, τ_5 — места глагола-связки перед предикативом, τ_6 — места глагола-связки после предикатива, τ_7 — иероглиф согласования с существительным в роде, τ_8 — иероглиф места определительной частицы после определения. Тогда предложение на хиндустани *сӯрья пӣт тара̄ хэ̄й* имеет входную иероглифику $\sigma_1\varphi_2 + \sigma_3\varphi_2\tau_1 + \sigma_2\tau_2 + \varphi_1\varphi_8\tau_6$, а его русский перевод *солнце — желтая звезда* — выходную иероглифику $\sigma_1\varphi_4 + \sigma_3\varphi_3\tau_1 + \sigma_2\tau_7\varphi_3\tau_2$, и, следовательно, при переводе с хиндустани на русский (и обратно) мера конгруэнтности преобразования в данном случае $K = 100 \times 10 : 19^2 = 2{,}8$. Рассмотрев иероглифику соответствующего китайского предложения *tàijáŋ ʂɨ̀ huáŋ-dè xīŋ* [2] — $\sigma_1 + \sigma_3\tau_1 + \sigma_2\varphi_7\tau_2 + \varphi_1\tau_5$, английского *the sun is a yellow star* — $\sigma_1 + \varphi_5\tau_4 + \sigma_3\tau_1 + \sigma_2\tau_2 + \varphi_6\tau_4 + \varphi_1\varphi_8\tau_5$ и французского *le soleil est une étoile jaune* — $\sigma_1 + \varphi_5\tau_7\tau_4 + \sigma_3\tau_1 + \varphi_6\tau_7\tau_4 + \sigma_2\tau_3 + \varphi_1\varphi_8\tau_5$, мы можем видеть, что при переводе с хиндустани на китайский, на английский или на французский мера конгруэнтности для данного случая соответственно равна $K = 3{,}3$; 2,9; 2,1, а при переводе с русского на китайский, английский или французский равна соответственно $K = 3{,}5$; 2,3; 1,5.

Отсюда мы получаем удельную конгруэнтность хиндустанского предложения $K_{ср.}$ равную $(2{,}8 + 3{,}3 + 2{,}9 + 2{,}1) : 4 = 2{,}8$, и удельную конгруэнтность русского предложения $K_{ср.} = 2{,}6$. Исходя из того, что при китайско-английском переводе $K = 3{,}2$, при китайско-французском $K = 3{,}0$ и при французско-английском $K = 2{,}3$, мы получаем удельную конгруэнтность китайского предложения $K_{ср.} = 3{,}2$, английского $K_{ср.} = 2{,}9$ и французского $K_{ср.} = 2{,}2$.

Высокая удельная конгруэнтность китайского не случайна: грамматическая структура этого языка имеет наименьшее число избыточных иероглифов, будучи всего ближе к «логическому синтаксису»; во французском же языке сочетаются избыточные при ряде переводов иероглифы грамматических родов и артиклей. Заметим сразу же, что картина получится иной, когда мы выйдем за пределы этой простой фразы, где мы считали полной конгруэнтность семантических иероглифов. В более широкой и разнообразной семантической иероглифике именно китайский язык обнаруживает наименьшую конгруэнтность — в силу того, что исторические традиции словообразования в этом языке были несходными с традициями европейских языков (что касается такого языка, как индонезийский, то он в этом отношении занимает промежуточную позицию, поскольку в нем оригинальные традиции переплетены с европейскими). Таким образом, общая оптимальная конгруэнтность по сумме семантических, формальных и тектонических иероглифов не является свойством какого-либо из живых языков.

При достаточно большом числе языков вступает в действие новый фактор: количество алгоритмов перевода возрастает быстрее числа языков. Для n языков нужно $n^2 - n$ программ. Очевидно, что введение языка-посредника сократит число алгоритмов до $2n$; при 100 языках выигрыш будет 50-кратным. В общих чертах об этом уже говорилось в статье П. С. Кузнецова, А. А. Ляпунова и А. А. Реформатского, а также в статьях Уивера, Бут и Локке [3].

---

[1] Вторая степень в знаменателе объясняется тем, что при одной и той же пропорции конгруэнтных иероглифов (допустим, 6 из 8 и 12 из 16) машина быстрее выполнит то преобразование, в котором иероглифический комплекс меньше. Иначе говоря, скорость работы машины есть функция двух факторов: однородности преобразования и его линейного размера. Умножение данных величин и дает нашу формулу.

[2] Этот и последующие примеры даются в новом алфавите.

[3] См.: П. С. Кузнецов, А. А. Ляпунов и А. А. Реформатский, Основные проблемы машинного перевода, ВЯ, 1956, № 5, стр. 109—110; W. Weaver, Translation, «Machine translation of languages», New York — London, 1955, стр. 23; A. D. Booth, W. N. Locke, Historical introduction, там же, стр. 7.

Весьма существенным представляется и тот выигрыш, который может быть получен, если язык-посредник будет обладать максимальной удельной конгруэнтностью по всем трем категориям иероглифов. Это может быть достигнуто путем устранения из языка-посредника элементов, иероглифы которых в большинстве преобразований являются неконгруэнтными [1], и путем сохранения элементов, иероглифы которых конгруэнтны в большинстве случаев перевода (иероглифика трех основных глагольных времен, грамматического числа существительных, степеней сравнения прилагательных и наречий, постановки подлежащего перед сказуемым [2] и др.). Необходимым свойством языка-посредника должно быть сведение к минимуму неконгруэнтной полисемии, осуществляемое путем логического анализа системы употребляемых понятий и способов ее языкового выражения.

Так мы приходим к проблеме структурального обобщения грамматических категорий наиболее распространенных языков мира и разработки международного терминологического кода с последующим созданием на этой базе языка-посредника машинного перевода. Очевидно, что язык-посредник должен быть создан отнюдь не на основе модернизованной латыни, а с учетом в равной мере языков Востока и языков Запада. С точки зрения машинного перевода одинаково существенно как то, с какого языка, так и то, на какой язык делается больше всего переводов. Поэтому даже теперь, когда научная продукция стран Европы и Америки временно превышает еще научную продукцию стран Азии и Африки, решающим является то обстоятельство, что количество переводов на незападноевропейские языки уже не уступает числу переводов на западноевропейские языки (согласно недавно опубликованным данным ЮНЕСКО [3], СССР стоит на первом месте по числу переводов, тогда как США — лишь на одиннадцатом). Поэтому строить машинный язык-посредник на базе только западноевропейских языков было бы прежде всего неэкономичным: его удельная мера конгруэнтности была бы слишком низка.

Стремление к оптимальной экономичности структуры языка-посредника (т. е. к максимальной мере конгруэнтности) обязывает принимать во внимание два фактора: 1) характер иероглифики языков, участвующих в переводческих отношениях; 2) количество переводов, выполняемых для того или иного языка. Не требуется доказательств того, что учет или неучет какой-либо черты языка, участвующего, скажем, в 20% случаев перевода, в бо́льшей степени влияет на экономичность машины, чем учет или неучет какой-либо черты языка, участвующего лишь в 1% случаев перевода. Однако было бы наивностью делать из этого вывод, что таким способом хотя бы в малейшей степени затрагивается вопрос о значимости того или иного языка: речь идет всего-навсего о коэффициенте полезного действия машинной конструкции.

Исходя из всех изложенных соображений, нетрудно получить математическое выражение грамматических и лексических типов, наиболее конгруэнтных при переводе с n на $n-1$ языков:

$$M = \frac{\sum_{i,j=1}^{n} q_{i|j}\, m_i\, m_j}{\sum_{i,j=1}^{n} m_i\, m_j},$$

где q — коэффициент, равный 1 при конгруэнтности двух иероглифов, участвующих в преобразовании с языка i на язык j, и равный 0 при неконгруэнтности, а $m_1, m_2, \ldots m_i, m_j, \ldots, m_n$ — веса участвующих в переводе языков.

При наличии языка-посредника функция упрощается и принимает вид:

$$\mu = \frac{\sum_{i=1}^{n} q_i\, m_i}{\sum_{i=1}^{n} m_i}$$

Мезофункция μ позволяет с математической строгостью определить грамматические черты и словарь языка-посредника, наиболее выгодные для режима работы машины.

[1] Таковы артикли германских, романских и арабского языков, классификаторы языков банту, тхайские, вьетнамские, китайские, бирманские и индонезийские счетные слова, грамматический род в славянских, новоиндийских, романских и немецком языках, славянские и арабские глагольные виды, согласование времен в хиндустани, романских и германских языках и т. п.

[2] Следует иметь в виду, что машинный перевод разрабатывается в первую очередь для научно-технических текстов.

[3] См. «Сов. культура» 5 II 57, стр. 4.

Продемонстрируем это на двух примерах. Предварительно припишем двадцати шести наиболее распространенным языкам мира веса, пропорциональные числу людей, которым эти языки практически известны из повседневной жизни: а) китайский — 24, б) хиндустани — 18, в) английский — 10, г) русский — 8, д) испанский — 5, е) японский — 4, ж) немецкий — 3, з) арабский — 3, и) индонезийский — 3, к) португальский — 2, л) французский — 2, м) суахили — 2, н) итальянский — 2, о) тюркские — 2, п) западнославянские — 2, р) иранские — 1, с) корейский — 1, т) хауса — 1, у) южнославянские — 1, ф) вьетнамский — 1, х) тхайский — 1, ц) бирманский — 1, ч) тагальский — 1, ш) румынский — 1, щ) скандинавские — 1, ъ) голландский — 1. Общая сумма весов — 101.

Рассмотрим, во-первых, выгоднее ли в языке-посреднике ставить прилагательное перед существительным, к которому оно относится, или после него. Имеем: а) *хӣn-dè ʂū*, б) *наи китāб*, в) *new book*, г) *новая книга*, д) *libro nuevo*, е) *атарасий хон*, ж) *neues Buch*, з) *китāб джадūд*, и) *buku jang baru*, к) *livro novo*, л) *livre nouveau*, м) *kitabi kipya*, н) *libro nuovo*, о) *yeni kitabI*, п) *nowa ksi̯ęga*, *nová kniha*, р) *кетаб-е ноў*, с) *сэ чхэк*, т) *sabo littafi*, у) *нова књига*, *нова книга*, ф) *sách mó'i*, х) *са-мут май*, ц) *Θu-ðŏ cāoў*, ч) *ang bagong aklat*, ш) *cartea nouă*, щ) *ny bok*, *ny bog*, ъ) *niew boek*. Ответ получаем точный и недвусмысленный; $\mu_1 = 0{,}77$; $\mu_2 = 0{,}23$; т. е. прилагательное в языке-посреднике должно стоять перед существительным.

При этом заметим, что уже по первым четырнадцати языкам мы могли бы получить соотношение приблизительно того же порядка $\mu_1 : \mu_2 = 69 : 19$. Поэтому, отыскивая в качестве второго примера выгодную для языка-посредника структуру простого распространенного предложения, ограничимся только частью взятой нами вначале совокупности языков: а) *ʐè-gè wùvèn-dè xìŋʐĭ wánqyán чуédìŋ-la tā-dè ʐŭjáu gūŋ'ǔŋ*, б) *васту-кӣ вишештāо-не ус-кā буниядū упаёг нишчит кия хэй*, в) *the properties of the object have fully determined its principal function*, г) *свойства предмета вполне определили его основную функцию*, д) *las propiedades del objeto han completamente determinado su función principal*, е) *моно-но сэйсйцу-ўа соно омонару сиё-о маттаку кэттэй-сйта*, ж) *die Eigenschaften des Gegenstandes haben seine Hauptfunktion vollständig bestimmt*, з) *ᶜäйänät χuçuçийāту-улı-māддāти ўäджибāт-χā 'äл-'äcāсийāтā тāмāмāн*, и) *sifat² dari benda itu telah menentukan djabatannja kepala setjara tjukup*, к) *as propriedades do objecto teem completamente destinado a sua funçāo principal*, л) *les propriétés de l'object ont complètement déterminé sa fonction principal*, м) *tabia za kitu imeyamurisha kabisa kazi kubwa yake*, н) *le particolarità del oggetto hanno completamente determinato la sua funzione principale*, о) *maddenin hususiyetleri esaslI tamamile tayin ettIlar*.

Получаем, что для взаимного расположения подлежащего и сказуемого $\mu_1 : \mu_2 = 85 : 3$, т. е. подлежащее должно стоять в языке-посреднике раньше сказуемого. Аналогичным образом приходим к тому, что сказуемое должно идти раньше прямого дополнения (61 : 27), что определения должны во всех трех случаях стоять перед определяемым (48 : 40, 69 : 19,78 : 10), что притяжательное местоимение должно быть поставлено до прилагательного (84 : 4) и наречие — до знаменательного глагола (64 : 24). Чтобы показать, как решается вопрос о формальных иероглифах, укажем на определительный формант к понятию «предмет» [такой формант целесообразно иметь в языке-посреднике (88 : 0)] и на артикль, который невыгодно иметь в языке-посреднике машинного перевода (30 : 58).

Эти два примера достаточно наглядно раскрывают методику применения мезофункции, т. е. методику численного решения вопроса о наиболее эффективной структуре языка-посредника.

П. Д. Андреев