ЛИНГВИСТИКА

ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ МАШИННОГО ПЕРЕВОДА: ДОПУСТИМЫЙ ВАРИАНТ

В.В. МАДОЯН,
К.А. АРУТЮНОВА
ЕГУАиС

Вопросы автоматического (машинного) перевода (АП, или МП), очень необходимого и пока реально трудно осуществляемого, все равно рано или поздно должны быть решены. Сегодня существуют три методики его реализации: Translation Memory (TM или Systems Mashine Translation - SMT), Rule-Based Mashine Translation (RBMT) и их комбинация. Несмотря на мощные ЭВМ и жесткие программы, проблема, на наш взгляд, не решается из-за отсутствия четкого лингвистического обеспечения. Синтаксическое и семантическое понимание текстов, морфологический анализ словоформ пока далеки от от необходимого. И такая ситуация складывается не потому, что лингвисты не в состоянии дать жесткое описание лингвистической системы, а потому, что АП требует недискретного описания, а современные лингвистические работы базируются на исследованиях дискретных.

Материал современных исследований привел нас к выводу о необходимости исчерпывающего СинСемП текста путем его анализа с применением математического моделирования с целью составления словаря однозначных словоформ. Поскольку полная систематизация материала оказывается невозможной, ниже анализ осуществлен с привлечением методики Translation Memory.

Выводы работы базируются на исследовании текстов длиной в 10 тысяч знаков, но само исследование демонстрируется на небольшом по объему материале – с его исчерпывающим описанием.

Если взять "Руководство по эксплуатации автомашин BMV", которое и использовано в нашей

работе в качестве материала исследования, и просистематизировать поверхностно-синтаксические схемы его предложений, как это делается в методике Translation Memory, то получим двенадцать схем, из которых глубинно-синтаксическими окажутся только четыре.

В то же время, не пренебрегая методом Translation Memory, ниже мы используем только одну поверхностно-синтаксическую структуру – с приписыванием каждой слово-форме те синсемантические и морфологические значения, которые, по нашим предложениям, должны быть в словаре словоформ для полного понимания текста, если он уже прошел информационный анализ и получил тематическую классификацию. Такой анализ мы называем исчерпывающим, поскольку, следуя математическому моделированию, приписываем каждой словоформе такие значения и в таком количестве, которые позволяют эту словоформу однозначно воспринять только на основании формальных (недискретных) данных.

Предложение, которое ниже мы переводим, является характерным для рассматриваемого технического текста и в своей структуре наиболее частотное, хотя и выбрано с долей вероятности, что вписывается в методику проведения подобных исследований. Итак, английский текст и его русский перевод:

Approx 300 mi/es (500 km) must elapse before the brake pads and rotors achieve the optimal pad-surface and wear patterns requiredfor trouble-free operation and long service life later on.

Следует пробежать около 300 миль (500 км), прежде чем тормозные колодки и колесные оси достигнут оптимальной поверхности соприкосновения и норм износа, требуемых для безотказной работы и длительного срока эксплуатации впоследствии.

Пронумеруем место каждого эквивалента – отдельной семантической единицы в обоих текстах. Это нужно для построения синтаксической схемы Т2 (языка, на который переводится текст; соответственно Т\ язык, с которого переводят) и для установления полного соответствия между компонентами оригинала и перевода.

При таком подходе – ввиду того что перевод с английского на русский – исключаем артикли, а такие единицы, как прежде чем и написанные через дефис классифицируем как цельные.

Получаем:

1	approx	около	3
2	300	300	4
3	miles	миль	5
4	500	500	6
5	km	км	7
6	must	следует	1
7	elapse	пробежать	2
8	before	прежде чем	8
9	(the) brake	тормозные	9
10	pads	колодки	10
11	and	и	11
12	rotors	оси	12
13	achieve	достигнут	13
14	(the) optimal	оптимальной	14
15	pad-surface	поверхности соприкосно-вения	15
16	and	и	16
17	wear	износа	18
18	patterns	норм	17
19	required for	требуемых для	19
20	trouble-free	безотказной	20
21	for operation	для работы	21
22	and	и	22
23	long	длительного	23
24	service	эксплуатации	24
25	life	срока	25
26	later on	впоследствии	26

Из рассматриваемого предложения получаем "русифицированный" вариант, т.е. располагаем слова так, как они располагаются в аналогичном русском предложении:

MUST ELAPSE APPROX 300 MILES (500 KM) BEFORE BRAKE PADS AND ROTORS ACHIEVE OPTIMAL PAD-SURFACE AND PATTERNS WEAR REQUIRED FOR TROUBLE-FREE FOR OPERATION AND LONG SERVICE...

Если составляющие данных структур представим в цифровой записи, получим Constr.XT1 (1-2... 26) = Constr. XT2 (3-4-5-6-7-1-2-8-9-10-11-12-13-14-15-16-18-17-19-20-21-22-23-24-25-26). Задача лингвистического обеспечения в данном примере заключается в том, чтобы словоформы были описаны на формальном уровне однозначно: так, чтобы программа могла их распознать и сопоставить. Эту задачу можно решить только одним способом: таким описанием словоформ, чтобы в сумме эти структуры не совпадали со структурами, состоящими из такого же количества единиц. Ввиду этого:

APPROX

Схема и соответствие

Constr. XT1 = Constr.XT2

Морф. 0 (неизменяемое)

Синт. D (определение),

DW - (определяемое слово) N -числительное (сочетание N+S)

Sem. ОКОЛО

Номерные соответствия в схемах: 1-3

300

Схема и соответствие

Constr. XT1 = Constr. XT2

Морф. N(N1) - числительное, им.п.

Синт. DW,

D - S (N2 pl) - существительное, род. п. мн. ч.

Sem. 300

Номерные соответствия в схемах: 2-4

MILES

Схема и соответствие

Constr. XT1 = Constr. XT2

Морф. S (N2pl.f) – существительное, род.п., мн.ч., ж.р.

Синт. DW,

D-N(Nl.pl)

Sem. МИЛЬ

Номерные соответствия в схемах: 3 -5.

Для экономии записи 500 km опускаем.

MUST

Схема и соответствие

Constr. XT1 = Constr. XT2

Морф. Vinper,3p,sing - глагол, безл., 3л., ед.ч.

Синт. PR.1 - сказуемое 1 (изменяемая форма составного глагольного сказуемого),

PR.2 - Inv.

Sem. СЛЕДУЕТ

Номерные соответствия в схемах: 6 -1.

ELAPSE

Схема и соответствие

Constr. XT1 = Constr. XT2

Морф. Inv.

Синт. PR.2,

PR - Vinper., 3p,sing.

Sem. ПРОБЕЖАТЬ

Номерные соответствия в схемах: 7-2.

BEFORE

Схема и соответствие

Constr. XT1 = Constr.XT2

Морф. 0

Синт. Conj.

Sem. ПРЕЖДЕ ЧЕМ

Номерные соответствия в схемах: 8-8.

BRAKE

Схема и соответствие

Constr. XT1 = Constr. XT2

Морф. ADjNl.pl - прилаг., мн.ч., им.п.

Синт. D

DW-S(N1Pl) КОЛОДКИ (SPH-ДВИЖЕНИЕ)

Sem. ТОРМОЗНЫЕ

Номерные соответствия в схемах: 9-9

PADS

Схема и соответствие

Constr. XT1 = Constr. XT2

Морф. S (N1 pl) - сущ., мн.ч., им.п.

Синт. DW - Subj - определяемое, подлежащее

D-AdjN1pl ТОРМОЗНЫЕ SPH ДВИЖЕНИЕ

PR - V3p.pl

Sem. КОЛОДКИ

Номерные соответствия в схемах: 10-10.

Подобные статьи можно продолжить. Построены они с отражением грамматики не английского, а русского языка. Казалось бы, грамматическое описание при наличии формы эквивалента на русском языке можно опустить, однако грамматическое описание – это путь к соединению методов ТМ и RBMT. Если программа в текстовой базе Т2 находит аналогичное по структуре предложение, которое отличается от переводимого только лексически, происходят лексические замены с учетом грамматических

признаков заменяемых слов. Если какие-то признаки не совпадают, происходит замена этих признаков и признаков согласуемых с данной формой слов. Например, если подлежащее переводимого предложения в ед.ч., а контрольного (базового) во мн.ч., такое же число принимают определяемые слова и сказуемое, поэтому в описании они фиксируются в одной статье (ср., например, описание pads).

Грамматические значения и номер эквивалента в структуре предложения позволяют производить однозначное определение формы слова, согласуемых с нею форм и устанавливать ее точное место в предложении. Вместе с тем такое восприятие текста свидетельствует о том, что у словоформ с точки зрения синтаксической функции могут быть иные постоянные и переменные признаки. Так, если с точки зрения классической грамматики для субстантивов падеж – переменная категория, то в предлагаемом варианте падеж – признак постоянный, поскольку в схеме любое существительное должно занимать данное место только в данном падеже (хотя и в ед., и во мн. ч.). Такие морфологические категории определяют сильные синтаксические связи в словосочетании (они указываются в словарной статье). А "чем больше связей в тексте окажутся семантически сильными, тем выше адекватность семантического анализа текста, так как он подтверждает семантическую связанность текста. Эта установка диктует требования к словарному описанию слов: они долж-ны быть достаточно общими, чтобы каждое новое употребление слова не объявлялось новым значением, а каждая связь не оказывалась новой, не имеющей отношения к уже установленным" [Леонтьева Н.Н., АПТ, с Л 22]. В то же время, если обратить внимание на произведенную нами выше семантизацию, точно фиксированными оказываются не только прямые и переносные значения слов, но и контекстуальные. Так, слово rotor в словарях может иметь значения "привод", "винт", "рабочий винт вертолета". Значение "ось", "ось колеса" выводится из контекста, поскольку тормозные колодки соприкасаются с колесом –"вращающейся частью". Слово life ни в одном словаре не фигурирует как "срок", однако в рус-

ском языке эксплуатация может иметь срок, а не жизнь. Значение слова устанавливается методом математического анализа, т.е. с привлечением всех факторов, которые влияют на выбор семантики, почему и предлоги приписаны как глаголу, так и существительному (ср.: required for и for operation).

Предлагаемое описание в определенной степени соединяет в себе ранний структурализм и поздний субстанционализм, с одной стороны, и дискретные описания (дескриптивной и функциональной лингвистики) с недискретными.

Как показано в работе, более эффективным является приписывание грамматических значений единицы речи Ti эквиваленту T2 и наоборот, поскольку важна грамматическая правильность перевода. По этой же причине значение артикля (названного нулевым) следует приписывать русскому языку (при переводе с русского на английский). Именно в виду этого нужна полная матрица грамматических значений обоих языков.

Достаточно вариативными оказываются не только члены просмотренных предложений, но и сама их структура, если ее рассматривать с точки зрения свертывания и развертывания через несколько шагов [Мадоян, Есаджанян]. Можно сказать, что синтаксис следует дополнить исследованием развертывания (свертывания) структур предложений, что дает новые сведения об их синтаксисе.

В современной лингвистике принято, что "темой считается семантически исходная часть предложения, т.е. предмет сообщения или то, о чем сообщается. Ремой называется то, что сообщается. Можно сказать, что рема и есть основное содержание предложения и тем самым его коммуникативный центр. Эти термины означают то же самое, что и коммуникативный субъект и коммуникативный предикат" [Кронгауз, с.207]. Такое тема-рематическое понимание сущности предложения, берущее свое начало с исследований Пражской лингвистической школы, практически повторяется без изменений и дополнений.

Рассмотрим в этом плане анализируемое нами предложение Approx 300 miles (500 km) must elapse before the brake pads and rotors achieve the optimal pad-

surface and wear patterns required for trouble-free operation and long service life later on.

Если структуру Const.X рассматривать с точки зрения синтаксической функции (синтаксической самостоятельности) ее частей, она окажется разбитой на несколько самостоятельных структур:

Approx 300 miles (500 km) must elapse

before the brake pads

and rotors achieve the optimal pad-surface

and wear patterns required for trouble-free operation

and long service life later on.

Самостоятельность частей проявляется в том, что любая из них может быть опущена с сохранением предикативности предложения и любая из них может в том же составе функционировать в другом предложении. Часть выделенных фрагментов может быть употреблена самостоятельно (можно говорить о самостоятельном употреблении всех частей, если убрать союзы). Для предложения в целом, как и первого фрагмента, выделение темы и ремы достаточно затруднительно. Здесь нужна не тема, а целая пресуппозиция. Таким образом, если отталкиваться не от тема-рематического понимания синтаксиса предложения, а от самостоятельности его частей, данная структура может свертываться вплоть до одной части, а одна часть может развертываться в большую структуру. Если следовать методике RBMT, можно разработать приемы выделения самостоятельных фрагментов текста, затем их классифицировать и установить порядок их следования друг за другом. Аналогичные операции в современной лингвистике применяются, например, при установлении правил согласования времен в английском языке. Это будут вполне дискретные описания, равные по значению недискретным (в плане применения в МП).

Правда, схемы могут различаться самыми разными признаками, например, часть из них может содержать актанты, часть - нет. Часть предложений может содержать опущенные члены и т.д.

При предлагаемом подходе меняется и программа МП. Она резко упрощается. Прежде всего меняется порядок анализа и синтеза. Анализ начинается с синтаксиса и кончается морфо-

логией возможных аналогов. Синтез также начинается с синтаксиса и кончается морфологией, что позволяет осуществить систематизацию материала (соединить методику ТМ с RBMT). Машина находит в базе текстов аналогичное предложение, отмечает несовпадающие словоформы и через словарь словоформ и словосочетаний производит их замену. При систематизации структур работа программы несколько усложняется, зато резко уменьшается база памяти. Уменьшается она и при систематизации частей предложения.

При таком подходе роль синтаксического анализа заключается в поиске аналога ПСС, а роль синтаксического синтеза завершается его русификацией.

По нашим предположениям и предлагаемому лингвистическому обеспечению, когда подается команда TRANS (ПЕРЕВОД), программа ищет допустимый аналог. После обнаружения возможного аналога, она расставляет составляющие английского предложения в том порядке, в котором они должны стоять в русском.

Предположим, подается предложение Over 200 miles (350 km) must elapse before the brake pads and rotors achieve the optimal pad-surface and wear patterns required for best operation. Машна находит рассмотренное выше предложение как контрольное (базовое), меняет не совпадающие словоформы, затем русифицирует синтаксис. Получаем: must elapse over 200 miles (350 km) before brake pads and rotors achieve optimal pad-surface and patterns wear required for the best operation.

Прежде всего через один шаг переводится аналогичная часть, затем начинается работа с неаналогами. С контрольным предложением не совпадают слова over, best. Слово over в конструкции Prepos + N (Numeralis) может иметь толко значение "свыше". Принимая нужную морфологическую форму, оно выступает как СВЫШЕ и занимает выделенное ему место.

Слово best в позиции перед существительным может иметь только значение "лучший". Поскольку этому слову задана форма МОРФ N2singf (род. п. ед.ч. ж.р.), best переводится как "лучшей".

Так как проблема выбора формы в лингвистическом обес-

печении уже решена, необходимо математическое описание семантики слова. Здесь может быть привлечена как сфера действия лингвистического знака, так и семантика сочетающихся слов слева и справа.

Если программа не находит в словаре то или иное слово, последнее остается на постранслящию, а пользователь может включить его в свой словарь, дабы в дальнейшем необходимости пострансляции не было вообще.

Что касается системного описания поверхностно-синтаксических структур и их систематизации, то это тема отдельного большого сочинения.

ЛИТЕРАТУРА

1. **Кронгауз М.Л.** Семантика. М., Academia, 2005.

2. **Леонтьева Н.Н.** Автоматическое понимание текстов. Системы. Модели.Ресурсы. М., Academia, 2006.

3. **Мадоян В.В., Есаджанян Б.М.** Русское предложение во взаимно порождающей классификации (предварительные замечания). Сб. "Русистика в СНГ", СПб., 2003.