ЛИНГВИСТИЧЕСКИЕ АСПЕКТЫ СИСТЕМ МАШИННОГО ПЕРЕВОДА

Г. Г. Бабалова, заведующий кафедрой иностранных языков Омского юридического института, кандидат филологических наук, доцент

Машинный перевод (далее - МП) есть выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке; причем преобразование основано не на семантическом анализе текста, как это делает переводчик-человек, а на анализе морфологии и синтаксиса. Это обстоятельство является основным источником проблем машинного перевода.

Отвлекаясь от технической реализации на чисто лингвистические аспекты проектирования систем машинного перевода, следует отметить следующее:

1. Каждая единица перевода может иметь несколько абстрактных представлений.

2. Необходимо предвидеть появление других неоднозначностей при движении по тексту от найденной, а также учесть в максимальной степени все другие возможные неоднозначности, которые могут возникнуть в результате принятого переводческого решения.

3. Чистый формализм не может служить основой анализа.

4. К формализму должны быть присоединены декларативные или процедурные лингвистические знания, причем способ присоединения таких знаний неясен.

Следует указать на некоторые проблемы, связанные не с конкретной моделью перевода, а с самим языком как объектом, принимающимся системой на входе, и объектом-результатом, то есть переведенным текстом.

Омоформия. Совпадение форм различных слов в написании и, возможно, в произношении, безусловно, является проблемой. В русском языке омоформия особенно распространена среди глаголов. Например, форма пошли может быть как формой прош. вр. от пойти, так и формой повелит. накл. от послать или пошлить; в предложном сочетании о тесте существительное может быть формой слов тест, тесть и тесто. Полные омонимы в сравнении с омоформами в русском языке редки (ключ, коса и др.). Омоформия вносит неоднозначность на этапе морфологического разбора. Данный вид неоднозначности в большинстве случаев устраняется при синтаксическом анализе в ходе отбрасывания недопустимых синтаксических конструкций.

Неоднозначность слов. Омонимия, омография и другие проявления различных значений одного и того же слова являются серьезной проблемой для программ машинного перевода. Например, слово разряд может быть спортивным, электрическим, двоичным и так далее. Видя данное слово в контексте, например: «мощный разряд вызвал ионизацию воздуха...»; или «перенос из старшего разряда приводит к переполнению...», можно понять, что в первом случае речь идёт об электрическом разряде, а во втором - о двоичном. Но даже в тексте не сказано явно, о каком разряде идёт речь.

Идеальным вариантом для переводчика является ситуация, когда каждая лексема имеет одно значение. Для реальных текстов, как правило, это невозможно. Неоднозначность является универсальным понятием для нескольких языковых уровней. Когда слово имеет более одного значения, говорят о лексической неоднозначности. Когда фраза (предложение) может иметь более одной структуры, говорят о структурной неоднозначности.

Неоднозначность есть неотъемлемая часть любого естественного языка. На практике количество вариантов перевода предложения значительно больше простой суммы вариантов переводов, в которых используется другое значение каждого его слова.

Так, предложение из 1О слов (каждое из которых имеет 2 варианта перевода) будет иметь 2^{1O} = 1О24 варианта анализа. Для реального случая таких вариантов будет гораздо больше. Рассмотрим следующие предложения:

You must not use abrasive cleaners.

The use of abrasive cleaners is not recommended.

В первом случае слово use выступает в качестве глагола, во втором - существительного. Традиционным подходом к решению проблемы является применение базы данных грамматических правил для конкретного языка. Так, в английском

языке невозможна грамматическая последовательность The + глагол + предложная группа, следовательно, перевод выполняется с помощью существительного.

Вполне очевидно, что, несмотря на кажущуюся формальность английской грамматики, такой подход может быть применен лишь в простых случаях. Можно дать системе МП информацию о грамматике языка в форме шаблонов, представленных с помощью регулярных выражений или каким-либо другим способом, что позволит фильтровать некоторые ошибочные варианты перевода. Однако эта информация не позволит установить значение во всех случаях неоднозначности. Одна и та же лексическая единица может иметь несколько значений даже для одной части речи, например, слово button. Во-первых, оно может быть как глаголом, так и существительным. Во-вторых, как существительное оно имеет два значения: «пуговица» и «кнопка».

Фактически вооружение системы одними лишь знаниями о синтаксисе может привести к потере эффективности, поскольку применение грамматики может порождать несколько различных вариантов анализа предложения в зависимости от примененного правила. Это может привести к возникновению огромного количества вариантов анализа для одного предложения.

Лексические и структурные несоответствия. На возникновение проблем лексических и структурных несоответствий могут повлиять два фактора: различия в способах представления мира в тех концепциях, которые выбираются носителем языка для выражения одиночных понятий; различия в структурах, которые используются носителями разных языков для выражения одной и той же мысли или, наоборот, одной структуры для мыслей различных.

Помимо этого, возникает так называемая проблема лексических дыр - случаев, когда в одном языке для выражения мысли используется целая фраза, в то время как в другом - одно единственное слово. Например, словосочетание affiliated societies переводится как «филиалы», а термин ambidextrous - фразой «свободно владеющий обеими руками». Проблемы, вызываемые лексическими дырами, похожи на проблемы, связанные с переводом идиом: в обоих случаях существует несоответствие в структуре переводимых конструкций.

Перевод устойчивых словосочетаний. Слова, входящие в состав устойчивых словосочетаний, зачастую имеют смысл в той или иной мере отличный от того, который они имеют, когда употребляются свободно. Поэтому такие сочетания следует переводить целиком, а не пословно. Для этого необходим словарь словосочетаний. При составлении словаря следует учитывать, что по синтаксической структуре словосочетания могут быть весьма разнообразны: яблоко раздора, камень преткновения, козёл отпущения, молодой человек, чёрный хлеб, железная дорога, скорый поезд, Российская Федерация, Соединённое Королевство, бить баклуши, валять дурака, задать стрекача, в зависимости от, в связи с, по прошествии и т.п. Встречаются и более сложные обороты, даже целые предложения, являющиеся известными высказываниями, пословицами и поговорками. Относительно последних следует иметь в виду, что они имеют тенденцию воспроизводиться не полностью (кто старое помянет...) и перифразироваться.

Представляется возможным предложить следующие практические критерии определения необходимости включения того или иного словосочетания в словарь: данное сочетание встречается достаточно часто (в речи, литературе, для специализированных систем МП - в текстах рассматриваемой предметной области); пословный перевод словосочетания на какой-либо язык неадекватен (пословицы, идиомы). Возможно, объем фразеологического запаса естественного языка окажется сравнимым с объемом его лексического словаря, а то и превзойдет его. Поэтому актуальной становится задача автоматического построения фразеологического словаря.

Идиомы. Идиома представляет собой сложившийся в языке, обычно эмоционально окрашенный оборот речи[1]. Особенностью идиом является то, что их общий смысл не мотивирован значением составляющих элементов и в общем случае не может быть выведен из них. Единицы, входящие в состав идиомы, полностью утратили семантическую самостоятельность и, следовательно, своими значениями не объясняют смысла всего оборота в целом. Это делает невозможным применение обычных схем и методов перевода. Таким образом, системе МП приходится работать с идиомами как с отдельными самостоятельными единицами.

Перевод идиом схож со случаем лексических дыр. Отличие между ними в том, что проблемы с лексическими дырами обычно возникают при переводе лексемы на язык, в котором ее значение выражается целой фразой. Проблемы с идиомами возникают при переводе с языка, в котором имеется идиома, на язык, в котором она выражается лексемой.

Существуют два традиционных подхода к решению проблемы идиом. Первый заключается в том, чтобы представить идиому как единицу, зафиксированную в словаре языка. При этом описываются морфологические правила для выделения этих единиц перед выполнением синтаксического анализа. Второй подход заключается в обработке идиом с помощью специальных правил, которые изменяют исходную идиоматическую структуру на соответствующую целевую структуру. Этот подход применим только в Transfer-системах и LK-системах. Еще одна проблема, связанная с идиомами, заключается в том, что зачастую возможна

либо идиоматическая, либо точная, дословная интерпретация конструкции.

Следует учитывать также, что для разбора идиом необходимы специальные правила, которые будут дополнять правила для обычных лексем и конструкций. Идиомы не всегда находятся в одной и той же форме, и количество форм не ограничено вариантами склонения, что является типичным для обычных слов. Таким образом, существует проблема выделения идиом в предложении. Она может проявляться не со всеми идиомами. Некоторые из них всегда присутствуют в предложении в одной форме, например устойчивые выражения или вводные конструкции типа in fact, in view. Иногда изменяется время (глаголы), лицо, число (существительные) у лексем, составляющих идиому.

Тематические признаки. Значение слова может зависеть от области знаний, к которой принадлежит переводимый текст. Общеизвестно, что каждая область знания имеет свой характерный лексикон. Если снабдить каждую словарную статью перечнем «тем», в которых данное слово может употребляться, и связать его возможные переводы с этими темами, то неоднозначность устраняется выбором того из вариантов перевода, который наиболее соответствует тематике текста.

В таком случае встает проблема определения тематики текста. Программа сама сможет определить ее, предварительно найдя в тексте слова, тематика которых достаточно однозначна. Например, слово депутат едва ли можно связать с чем-либо, кроме политики, а встретив в тексте термины морфология, орфография и т.п., программа может «догадаться», что в нем говорится о лингвистике, и слово согласный уже будет переводить как consonant, а не agreeabЫe. Наиболее трудной задачей при реализации этого подхода является разработка системы отнесенности лексем и словосочетаний к определенной области знания. Для этой цели должны быть решены вопросы детальности разбиения лексики (нейтральная лексика, язык художественной литературы, научной литературы, публицистики, разговорный) и ее системной организации, возможно, с помощью метода компонентного анализа. Рациональной представляется иерархическая организация лексики:

1. Общеупотребительная лексика.

2. Научная:

– естественнонаучная: математическая, физическая, компьютерная и др.;

– гуманитарная: педагогическая, юридическая, историческая и др.

3. Публицистическая.

4. Деловая.

5. Разговорная и т.д.

Учет особенностей словообразования. Русскому языку присуще разнообразие словообразовательных средств. В английском языке словообразование играет меньшую роль. Многие слова имеют похожую словообразовательную структуру, например: ежегодный, ежемесячный, ежечасный, ежеминутный, ежесекундный.

Схожесть структуры во многих случаях означает и схожесть значения, а нередко - и перевода. Особенно это справедливо для составных слов, например: черно-белый - black and white; красно-коричневый - red and brown; серо-зеленый grey and green; седовласый - grey-haired; востроглазый - keen-eyed; длинноногий - long-legged. Перевод таких слов достаточно прост, если знать перевод их составляющих. Поэтому вместо того, чтобы заносить каждый отдельный случай в базу данных системы, удобнее иметь возможность описать их одним правилом: перевод (X + «о-» + Y + «ый») = перевод (X) + «and» + перевод (Y), если X и Y суть основы прилагательных. Перевод (X + «о» + Y + «ый») = перевод (X) + «-» + перевод (Y) + «ed», если X - основа прилагательного, а Y - основа существительного[2].

Наличие в системе МП информации о словообразовании дает следующие плюсы: уменьшение размера лексико-морфологического словаря по сравнению с подходом, когда для каждого слова хранится его полная основа; возможность задания общего шаблона перевода для слов с похожей словообразовательной структурой и отсюда - уменьшение размера двуязычного словаря. Недостаток этого подхода заключается в том, что он требует более кропотливого труда при составлении словаря. Однако процесс морфемного членения слов можно частично автоматизировать.

Словообразовательную информацию можно представить следующим образом. Сначала необходимо составить словари морфем: корней, приставок и суффиксов. Собственно словарь может представлять собой набор таблиц, в каждой из которых будут сведены слова с одинаковой словообразовательной структурой. Например, существительные, образованные суффиксальным способом от немотивированных основ (корней), могут быть объединены в одну таблицу. Основы слов в этой таблице могут быть представлены двумя полями: «Корень» - индекс записи в таблице «Корни» и «Суффикс» - аналогичный индекс записи в таблице «Суффиксы». Слова с каким-либо особо продуктивным суффиксом можно выделить в отдельную таблицу, тогда о наличии этого суффикса у слова будет говорить его принадлежность к этой таблице.

Вследствие естественных ограничений, а также по экономическим причинам системы МП применяются только для межъязыкового перевода. Учитывая эти недостатки, мы не можем говорить о системе МП как об универсальном переводчике. Пока нет возможности алгоритмически воспроиз-

водить творческую деятельность переводчика, однако имеется возможность исследования формализации его деятельности по отношению к тексту. Иными словами, можно попытаться с помощью ЭВМ сделать анализ текста и выявить те его особенности, на основе которых переводчик определяет стилистическую окраску, выбирает лексические единицы из ряда синонимов, омонимов и т.д. Наиболее трудной задачей является разработка системы отнесенности лексем и словосочетаний к определенной области знания. Разработка компонентной структуры терминов предметной области может помочь в решении этой задачи[3].

[1] См.: Языкознание. Большой энциклопедический словарь / гл. ред. В. Н. Ярцева. 2-е изд. - М.: Большая Российская энциклопедия, 1998.

[2] См.: Слепов Н. Н. Некоторые идеи по архитектуре системы, структурам данных и алгоритмам. Проект «Машинный перевод» // http://mt.slova.tk.

[3] См.: Бабалова Г. Г. Лингвистические аспекты информатики (терминология и лексикография). - Омск: изд-во ОмГПУ, 2ОО4. - С. 84.

ИСПОЛЬЗОВАНИЕ МЕЖПРЕДМЕТНЫХ СВЯЗЕЙ В ПРОЦЕССЕ ПРЕПОДАВАНИЯ МАТЕМАТИКИ И ИНФОРМАТИКИ В ЮРИДИЧЕСКОМ ВУЗЕ

М. А. Екимова, старший преподаватель кафедры правовой информатики Омского юридического института, кандидат педагогических наук

Межпредметные связи в обучении являются отражением интеграционных процессов, происходящих сегодня в науке и в жизни общества. Они играют важную роль в улучшении научно-теоретической и практической подготовки студентов. Их ведущая функция состоит в формировании у студентов целостных знаний. Осуществляя межпредметные связи различных учебных дисциплин, преподаватель учит студентов видеть проблему с различных сторон, комплексно, и применять взаимосвязанные знания и умения для ее решения.

Межпредметные связи определяют по-разному: как принцип обучения; как условия воспитывающего и развивающего обучения; как дидактические условия, обеспечивающие качественное усвоение знаний студентами, развитие их мышления, творческих способностей. Но одно несомненно: их значение в педагогическом процессе неуклонно возрастает.

На сегодня в вузах не в полной мере реализуются принципы преемственности и системности содержания образования, нарушены межпредметные связи[1] . Обучение в них построено таким образом, что в редких случаях два преподавателя, которые ведут различные учебные дисциплины, сотрудничают на одном занятии, не намного чаще можно видеть интеграцию родственных дисциплин.

Современное общество характеризуется высокой степенью информатизации профессиональной деятельности специалистов, информация постоянно меняется, увеличивается ее объем, развиваются информационные технологии. А значит, для обеспечения эффективности обучения необходимо осуществлять межпредметные связи информатики с другими учебными дисциплинами (юриспруденцией, математикой[2], статистикой, экономикой[3] и др.), использовать средства вычислительной техники в процессе обучения различным дисциплинам.

Учебная дисциплина «Информатика и математика», изучаемая в юридических вузах, состоит из двух разделов: информатика и математика. И информатика, и математика являются самостоятельными науками, поэтому было бы уместно не соединять их в одну учебную дисциплину. Но в то же время эти науки сильно связаны, значит, необходимо в процессе обучения осуществлять их межпредметные связи.

Заметим, что при преподавании учебной дисциплины «Информатика и математика» межпредметные связи уместно использовать при обобщающем повторении - это помогает студентам повторить определенные темы и закрепить навыки. Рекомендуется использовать математические задачи для закрепления навыков работы с электронными таблицами. С этой целью на кафедре правовой информатики ОмЮИ были разработаны и апробированы несколько заданий, которые студенты выполняют на практических занятиях по информатике.

Темы занятий:

1. Случайные величины.
2. Элементы математической статистики.
3. Линейная корреляция.

Цели занятий:

1. Научить студентов находить решения матема-