Pazhuhesh-e Zabanha-ye Khareji, No. 19, Special Issue, Russian, 2005, pp 43-58

Термины в автоматической обработке текстов

Юрий Николаевич Марчук

Профессор Московского государственного университета имени М.В. Ломоносова

e-mail: general@philol.msu.ru

Краткое содержание

Научно-технический прогресс в значительной мере зависит от распространения научно-технической информации. Важную ее часть составляют научно-технические переводы, объем которых в мире неуклонно возрастает в связи с расширением потребностей научно-технического общения. Тем не менее, исследования по искусственному интеллекту показали, что четкое раз навсегда заданное отделение исходных данных от алгоритма решения задачи не всегда достижимо, и, что главнее, не всегда целесообразно. Автоматическая обработка текстов естественного языка в практическом плане чаще всего ограничивается текстами определенных предметных полей, которые обладают признаками специальных подъязыков естественного языка. Автоматическая обработка текстов на естественном языке с необходимостью должна решать вопросы анализа и синтеза на основных языковых уровнях: морфологическом, синтаксическом, семантическом. Наш опыт показывает, что для правильного анализа и синтеза терминов в автоматизированных системах, необходимы специальные словари. Именно в таком случае можно надеяться на правильный и эффективный перевод или смысловой код рассматриваемого термина.

Ключевыеслова: машинный перевод, анализ текста, терминалогия, обработка текста,

Введение

Современная языковая ситуация характеризуется возросшим количеством терминов в текстах на естественных языках, подлежащих автоматической обработке. Изучением, анализом и регулированием терминопользования занимается наука терминология, точнее, терминоведение, поскольку слово «терминология» имеет также значение «система терминов конкретной предметной области» (например, медицины). Теперь уже практически нет таких языковедов, которые бы считали слова-термины всегда однозначными и не представляющими интереса для лингвистов. Терминоведение стало важной частью лингвистической науки (Авербух, 2004, стр.251) .

Роль правильного понимания термина чрезвычайно велика как в практической жизни, например, в законотворчестве и в применении законодательства, так и в теории и практике языкознания, не только прикладного, но и теоретического и сопоставительного. Термины составляют неотъемлемую и важную часть современной массовой коммуникации, особенности которой с точки зрения лингвистики и филологии очень подробно исследовал академик Ю.В. Рождественский (Рождественский, 2003, стр. 239).

Автоматическая обработка текстов естественного языка в практическом плане чаще всего ограничивается текстами определенных предметных полей, которые обладают признаками специальных подъязыков естественного языка. Так, можно

выделить подъязыки общенаучной лексики, подъязыки определенных областей науки и техники, причем среди последних могут быть как достаточно обширные подъязыки, например, подъязык текстов по машиностроению, так и сравнительно ограниченные, например, подъязык текстов по робототехнике в рамках более широкого подъязыка – текстов по машиностроению.

Какие проблемы с терминами возникают при автоматической обработке текстов?

Сначала мы уточним, что понимается под «автоматической обработкой текстов». В современной языковой ситуации распространены системы искусственного интеллекта, которые с помощью компьютерных программ анализируют содержание текста для того, чтобы поместить проанализированный текст или некоторое его воспроизведение в соответствующее место информационной базы. Разновидностями таких систем являются системы распознавания устной речи, обучающие системы, которые применяются для дистанционного обучения или в помощь преподавателю в классе, как , например, компьютерная система, помогающая преподавателю английского языка работать с учащимися, для которых родным языком является персидский (Rahimi, 2002, стр. 127), системы логического вывода, экспертные системы, которые моделируют поведение человека-эксперта при решении интеллектуальных задач, системы информационного поиска в базах данных и в базах знаний, системы машинного

перевода. Последние можно считать наиболее сложными и требующими весьма глубоких разработок в области моделирования интеллектуальной деятельности человека. Академик Ю.В.Рождественский считает системы машинного перевода центральными в проблеме искусственного интеллекта (Рождественский, 1987, стр.116).

Автоматическая обработка текстов на естественном языке с необходимостью должна решать вопросы анализа и синтеза на основных языковых уровнях: морфологическом, синтаксическом, семантическом. Наиболее в теоретическом и практическом отношении разработаны вопросы автоматического морфологического анализа. На морфологическом уровне существует наибольшее количество формальных признаков, которыми можно пользоваться для составления алгоритмов анализа и синтеза. Морфологический анализ входных словоформ является начальным этапом анализа, на данных морфологического анализа базируются все дальнейшие этапы алгоритмов автоматического анализа и синтеза текстов, практически во всех приложениях компьютерной лингвистики. Морфологический анализ достаточно хорошо разработан даже для перевода с европейских языков на восточные (Мосавимиангах, 2002, стр.151). Следующий уровень языковой структуры – синтаксический. Здесь гораздо меньше приемлемых теоретических решений, правильное определение синтаксической структуры требует весьма сложных исходных данных, тесно

связанных с семантикой. Абстрактные теории синтаксического анализа часто не дают приемлемых прикладных решений и не позволяют строить эффективные алгоритмы, поскольку в их (теорий) построениях авторы исходят из идеальных представлений о синтаксических зависимостях в языковом высказывании. В реальных текстах, не говоря уже об устной речи, мы часто встречаемся с такими построениями, которые понятны только в условиях экстралингвистического контекста. Очень часто реально встречающиеся в текстах синтаксические конструкции не удовлетворяют регулярным правилам их порождения и содержания, но тем не менее являются нормативными, входят в языковую норму. Поэтому многие ученые, занимавшиеся в течение долгого времени поисками универсальных алгоритмов синтаксического анализа, признают недостижимость всегда правильного автоматического синтаксического анализа на базе строгих формальных правил и абстрактных теорий. Очень часто эти теории ограничиваются узким кругом синтаксических явлений, что также недостаточно для большинства практических применений. Между тем системы автоматического синтаксического анализа, построенные на принципах использования модели переводных соответствий в рамках конкретных языковых пар, показывают достаточную эффективность (Валипур, 1998, стр. 211)

Остается семантика. Этот уровень самый сложный, он, с одной стороны, весьма тесно связан с синтаксисом. Напомним о

том, что школьный грамматический разбор предложений базируется на смысле, содержании предложения. «Что такое подлежащее? Это главный член предложения, то, что осуществляет действие. Что такое сказуемое? Это то, что делает подлежащее» и т.п. Семантический уровень предложения или высказывания, с другой стороны, тесно связан с экстралингвистической реальностью , а также и с содержанием всего текста в целом. В.А. Звегинцев приводит такой пример, характеризуя смысловое содержание разных вариантов описания (перефраз) , казалось бы, одной ситуации: «Волк был убит ударом ноги выскочившего из-за куста охотника»; «Охотник выскочил из-за куста и ударом ноги убил волка»; «Ударом ноги охотник убил волка, выскочив из-за куста» и т.п. Можно ли считать данные высказывания эквивалентными по смыслу? В.А. Звегинцев обоснованно считает, что нет. Каждый из этих вариантов имеет свой собственный смысл, зависящий от контекста, как лингвистического, так и экстралингвистического.

В последнее время Институтом языкознания Российской Академии наук проведен ряд исследований, посвященных детальному изучению понятия «смысл». На основе работ А.И. Новикова (Новиков, 2002, стр. 155-180) был поставлен эксперимент по различению смысла и содержания текста. Информантам были даны три текста: научный текст по химии, исторический (походы Александра Македонского) и художественный (описание разговора мужчины и женщины,

едущих в поезде и обменивающихся впечатлениями от дороги). По результатам опроса информантов выяснилось, что, по их мнению, смысл и содержание первого текста, научного, совпадают. Смысл и содержание исторического текста различаются: содержание состоит в описании походов, а смысл – показ завоевательной деятельности полководца. В третьем тексте смысл и содержание совершенно не совпадают: содержание – диалог о природе, а смысл – выяснение личных отношений.

Смысл и содержание тесно связаны, с одной стороны, с конструкцией предложения как суммарной составляющей отдельных его частей, так и с семантикой самой главной составляющей всякого высказывания – словом. Нет необходимости доказывать, что основной частью всякого речевого высказывания является слово. Даже в тех языках, где граница слов может быть не так четко определена, как в языках европейских, понятие «слова» существует. Примером может служить китайский язык, в котором исследователи отмечают не только слова, но и категорию частей речи. Слово, таким образом, является центральной единицей как языка, так и речи. Поэтому распознавание, анализ, идентификация роли слов в высказывании является основой всяких других видов анализа, в том числе и автоматического. Именно в уходе на уровень лексики и семантики слов видят современные исследователи выход в создании и развитии алгоритмов синтаксического и семантического анализа.

Одним из наиболее распространенных видов слов в современной лексике, как научно-технической и специальной, так и общеупотребительной, является слово-термин, т.е. слово, так или иначе обозначающее довольно точное понятие или явление. Здесь есть несколько важных проблем. Первой из них является отличие термина от общеупотребительного слова.

Существуют термины широких предметных областей, такие как *двигатель, разработка, машина, программа, инерция и т.п.* которые также безусловно являются и общеупотребительными словами. Например, в сочетаниях *двигатель прогресса, инерция мышления* и пр. эти слова следует отнести к общеупотребительным, в крайнем случае рассматривая их как метафору, однако нужно отметить, что их метафоричность в настоящее время практически стерлась. В то же время слово *двигатель* в текстах по автомобилестроению совершенно определенно означает конкретную часть автомобиля, отличающуюся от других частей и имеющую совершенно определенную структуру и функции. Действенным средством отличия терминов от не-терминов является контекстологический словарь, о котором далее пойдет речь. Второй проблемой является понятийная атрибуция термина, т.е. точное соотнесение термина и соответствующего понятия предметной области. Такой атрибуции мешает полисемия термина как слова естественного языка. Количество строго однозначных терминов всегда относительно небольшое даже для малых и

специфических предметных областей. Даже, например, в текстах по стоматологии или медицинской генетике (см, например (Оганесян, 2003, стр. 215) большое количество терминов широкой семантики. Так, в англо-русском словаре медицинской генетики встречаем слова: *translation* - заключительный этап реализации генетической информации, синтез полипептидных цепей рибосомами; *transmission* - передача наследственных качеств родителей потомству; *arm* - плечо хромосомы, и т.п.

Место термина в системе понятий данной предметной области может совершенно точно установить только специалист в данной области. Однако это не означает, что терминологические проблемы в текстах должен и может решать только специалист. Лингвист, имеющий дело с текстами, также не в меньшей степени может правильно решать проблемы правильного использования терминологии. Каким образом? С помощью учета использования термина в текстах, исследования его окружения.

В текстовой идентификации термина основную роль играют два фактора: определение (дефиниция) термина, которая дается в терминологическом словаре соответствующей предметной области и фактическое употребление термина в текстах. Эти два фактора безусловно могут использоваться лингвистом при решении вопроса об атрибуции термина.

Второй фактор в настоящее время является ведущим. Новейшие лингвистические исследования все более склоняются к

использованию текстов как основного исходного материала при описании и определении лингвистических единиц. Именно использование, или, до недавнего времени, дистрибуция термина являются главным фактором его атрибуции, или рекомендаций по его применению. Современные направления лингвистических исследований базируются на изучении определенных корпусов текстов для выявления лингвистических характеристик. Так называемая корпусная лингвистика, приемы которой были известны уже достаточно давно и применялись в прикладных целях, таких, как дистрибутивное описание лингвистических единиц, теперь все более четко формулируется как теоретическая дисциплина (Марчук, 2002, стр. 234). Интересно отметить, что даже такие лингвистические инструменты, как тезаурус, где первоначально слова группируются по дедуктивному теоретическому представлению о содержании понятий, в настоящее время создаются на базе исследования представительных корпусов текстов, а не на основе заранее заданных теоретических представлений. Теоретические концепции, в том числе и группировки лексических единиц, получаются индуктивным путем, а не задаются заранее, хотя и в этом последнем пути продолжаются некоторые ограниченные применениями исследования.

Разновидностью атрибуции термина является его перевод на другой язык. При переводе требуется не только правильно понять предметное содержание термина, но и найти для него точный

переводной эквивалент. Этот вопрос интересовал терминоведов и терминологов с самого начала становления терминоведения как научной и прикладной дисциплины (см., например, труды основоположника научного терминоведения Д.С. Лотте). Даже при установлении предметной соотнесенности не всегда возможно найти правильный переводной эквивалент. Так, мне приходилось во время работы в Камбодже решать следующую проблему: найти эквивалент русского термина *ледник* в кхмерском языке. Такое слово есть в языке кхмеров, но оно означает небольшое хранилище со льдом, нечто вроде домашнего холодильника. *Ледник* как ледовая река, поток льда, стекающий с гор, не укладывается в языковое сознание кхмеров, которые не видели никогда, например, снега, идущего в холодную погоду. В таких случаях, с помощью специалистов, для которых кхмерский язык является родным, приходится находить некоторое описательное определение.

Однако во всех других случаях изучение дистрибуции, или использования термина, в представительном массиве текстов – в случае перевода, в переведенных текстах – позволяет найти и установить правильный его перевод в зависимости от контекста.

Остановимся на чрезвычайно важном инструменте контекстного анализа – автоматическом контекстологическом словаре для перевода многозначных слов. Идея контекстологического словаря была опубликована мной еще в восьмидесятых годах прошлого века , а сам словарь для

машинного перевода многозначных слов с английского языка на русский был издан в 1976 году. Этот словарь явился плодом многолетних разработок большого научного коллектива исследователей в рамках проекта машинного перевода. Принцип контекстологического словаря – разрешение лексической многозначности конкретного слова (лексемы) в зависимости от ее употребления в исходном корпусе текстов. Употребление лексемы в исходном корпусе текстов фиксируется в словаре-конкордансе, составленном на исходном массиве. Этот массив должен быть достаточно представительным.

Словарь-конкорданс содержит перечень всех употреблений словоформ данной лексемы в текстах. В данном словаре становится возможным для исследователя обнаружить в текстах морфологические, синтаксические, семантические и лексические характеристики, которые служат детерминантами, т.е. по отдельности или в некоторой совокупности однозначно определяют значение многозначного слова. При этом значение приравнивается переводу на выходной язык, и словарь-конкорданс составляется по параллельным текстам – текстам на исходном (входном) языке и текстам переводов на выходной язык. В случае отсутствия детерминанты слову дается некоторый общий перевод, статистически и семантически обоснованный (Марчук, 2002). Использование контекстологического словаря в системах машинного перевода, построенных по модели переводных соответствий, показало его эффективность в переводе довольно

широких диапазонов текстов определенных предметных областей.

Критики концепции контекстологического словаря обычно указывают на трудоемкость его составления. Однако многолетняя практика машинного перевода показала, что другого пути разрешения лексической многозначности просто не существует. Попытки теоретически разрешать многозначность набором универсальных семантических множителей, лексическими функциями и другими подобными дедуктивными построениями не дали никаких практических результатов , хотя исследования в этом направлении ведутся уже много лет. Можно провести параллель между такими методами и попытками использовать универсальные искусственные языки типа эсперанто для машинного перевода сразу с нескольких языков. Многолетние исследования по проекту ЕВРОТРА для Европейского экономического сообщества, как известно, закончились ничем. Кроме того, все ныне действующие системы машинного перевода, имеющие практическое применение, используют тот или иной вариант модели переводных соответствий в конкретных языковых парах (Хроменков, 2000, стр. 210).

В последнее время А.Л.Семеновым обосновано использование контекстологического словаря как основного средства организации многоязычных политематических баз данных (Марчук, 2002, стр. 234). Определение термина, которое

приводится в таком словаре, сопровождается примерами использования термина в текстах. Такие примеры могут иллюстрировать использование данного термина как в одноязычных текстах, поясняя его значение в предметной области, так и в двуязычной ситуации, приводя примеры его перевода. Лексические детерминанты в таких случаях содержат термины, находящиеся с исходным в определенных иерархических зависимостях. Так, если заглавное слово представляет собой родовой термин, то лексические детерминанты, приводимые в контекстах, в основном являются видовыми терминами. В этом случае посредством лексических употреблений вскрывается и иллюстрируется иерархическая структура предметного поля данной отрасли. Именно таким образом построен словарь англо-русских – русско-английских терминов страхового дела С.В.Меркуловой (Меркулова, 2000, стр.134).

Заключение

Суммируя рассуждения настоящей статьи, можно утверждать, что для правильного анализа и синтеза терминов в автоматизированных системах, предназначенных для обработки естественно-языковых текстов , необходимы специальные словари, содержащие как дефиниции терминов, так и примеры их контекстного употребления – автоматические контекстологические словари. Именно в таком случае можно

надеяться на правильный и эффективный перевод или смысловой код рассматриваемого термина.

Литература

1- Авербух, К.Я., ***Общая теория термина***, Иваново, Изд-во Ивановского МГУ, 2004.

2- Валипур, Али Реза, ***Анализ и синтез глагольных форм и конструкций при машинном переводе с русского языка на персидский язык***, *Канд. дисс*, Москва, МГУ. 1998.

3- Довбыш, О.В., ***Англо-русский словарь терминов финансовой отчетности***, Самара, Парус-Принт, 2003.

4- ***Материалы к компьютерному тезаурусу лексики русского языка. Материалы конференции*** «Корпусная лингвистика и лингвистические базы данных». Санкт-Петербург, Изд-во Санкт-Петербургского Университета, 2002.

5- Марчук, Ю.Н., ***Основы компьютерной лингвистики***. Москва, Изд-во МГОУ, 2002.

6- Меркулова, С.В., ***Толковый словарь страховых терминов***. Dictionary of Insurance Terms. Москва, Изд-во МАИ, 2000.

7- Мосавимиангах, Тайеби, ***Морфологический анализ в системе англо-персидского машинного перевода***, Канд. дисс., Москва, Институт Языкознания РАН, 2002.

8- Мохаммади, М.Р., ***Система русско-персидского машинного перевода на основе переводных соответствий***. *Канд.дисс*, Москва, МГУ, 1998.

9- Новиков, А.И., ***Доминантность и транспозиция в процессе осмысления текста***, В кн: Проблемы прикладной лингвистики. Москва, Институт Языкознания, РАН, 2002.

10- Оганесян, М.В., ***Сопоставительно-переводческий анализ английской и русской медицинской терминологии по генетике***, *Канд. дисс.*, Москва, МГОУ, 2003.

11- Рождественский. Ю.В., ***Философия языка***, Культуроведение и дидактика. Москва, Грантъ, 2003.

12- Рождественский, Ю.В., Волков, А.А., Марчук, Ю.Н., ***Введение в прикладную филологию***. Москва, МГУ, 1987.

13- Хроменков, П. Н., ***Анализ и оценка эффективности современных систем машинного перевода***. *Канд.дисс.*, Москва, МПУ, 2000.

14- Эммарлу Рамезанали, ***Синтаксический анализ и синтез именных словосочетаний при машинном переводе с русского языка на персидский***. *Канд. дисс.*, Москва, МГУ, 1998.

15- Mehrak, Rahimi, ***Introduction to Teaching English as a Foreign Language***. Москва, Народный Учитель, 2002.