ISSN 0548-0027

Colinq 88

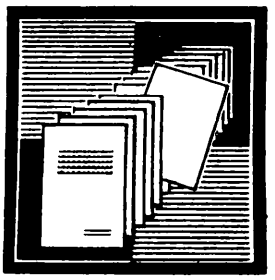
3'89

ВИНИТИ

НАУЧНО-ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

СЕРИЯ 2

ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ



СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 061.3:519(498)

О. С. Кулагина, Е. В. Падучева, Е. В. Рахилина

О КОНФЕРЕНЦИИ ПО ВЫЧИСЛИТЕЛЬНОЙ ЛИНГВИСТИКЕ В БУДАПЕШТЕ (COLING-88)

Дается обзор работ, представленных на XII Международной конференции по вычислительной лингвистике (COLING-88), состоявшейся в августе 1988 г. в Будапеште. Среди докладов конференции основное место занимали описания действующих систем автоматической обработки текстов на ЭВМ.

XII Международная конференция по вычислительной, или компьютерной лингвистике (COLING-88), организованная Международным комитетом по вычислительной лингвистике (ICCL) и обществом Дж. фон Нойман по вычислительным наукам (NJSZT) совместно с Институтом Венгерской Академии наук, проходила в Будапеште с 22 по 26 августа 1988 г. В ней приняли участие около 600 специалистов из 35 стран; наиболее крупными были делегации ФРГ, США, Японии, Швеции, Франции. От СССР приняли участие несколько десятков специалистов, четверо из них — с докладами.

На конференции было заслушано 135 докладов, проведено четыре дискуссии, были организованы демонстрации работы систем, выставка книг. Доклады были распределены по следующим 12 темам: Машинный перевод — 24 доклада; Формальные модели — 17 докладов; Анализ текстов — 15 докладов; Дискурс — 14 докладов; Семантика — 10 докладов; Синтаксис и морфология — 10 докладов; Понимание и представление знаний — 10 докладов; Лексика — 8 докладов; Анализ и синтез устной речи — 8 докладов; Обучение с помощью компьютеров — 7 докладов; Синтез текстов — 7 докладов; Математическое обеспечение — 5 докладов.

МАШИННЫЙ ПЕРЕВОД

Изложение разделено на две части. К первой отнесены в основном доклады, в которых центр тяжести — описание систем автоматической обработки текстов на ЭВМ; во второй части большее внимание уделяется чисто лингвистическим аспектам работы этих систем. Наибольшее число докладов было посвящено машинному переводу текста с одних языков на другие. Машинный перевод, начавший бурное развитие в пятидесятых годах, прошедший спад в конце шестидесятых — начале семидесятых годов, начал новый подъем в конце семидесятых годов и сейчас активно развивается, расширяя как сферу деятельности, так и свою географию. В последние годы начались работы по машинному переводу в Испании, Южной Корее, Индии, Индонезии, после довольно долгого перерыва возобновились работы в КНР.

Особенно заметна активизация работ по машинному переводу в последнее десятилетие в тех странах, для которых по тем или иным причинам перевод является насущной необходимостью. В первую очередь это — Япония, Канада, где два государственных языка, а также страны Европейского экономического сообщества, все документы которого должны переводиться на языки всех стран сообщества.

В настоящее время в мире имеется уже большое число практически действующих коммерческих систем. Они делятся на два типа. Одни — это системы, рассчитанные на ограниченную тематику и определенный вид входных текстов; другие — это системы без ограничений на вид входных текстов, обычно имеющие очень большие словари и дающие грубый перевод, требующий существенного постредактирования. Однако благодаря тому, что редактор работает за пультом ЭВМ, снабженной богатыми средствами редактирования, пара ⟨переводческая система + человек-редактор, располагающий автоматизированными средствами внесения редакторской правки⟩ работает в три-четыре раза быстрее, чем пара ⟨человек-переводчик + человек-редактор, работающие традиционными средствами⟩. Грубый машинный перевод полезен также в ситуациях, когда иными средствами получить перевод трудно (редкий язык, нет переводчиков), или тогда, когда пользователь может удовлетвориться общим представлением о содержании текста.

Наряду с системами, работающими в пакетном режиме, к которым человек подключается на заключительной стадии редактирования, все большее распространение получают системы, работающие в интерактивном режиме. Человек может участвовать в их работе

на самых разных этапах: на начальном, проводя в сущности предредактирование, на этапе анализа, преобразования или синтеза.

Преобладающий тип систем машинного перевода, разрабатываемых в настоящее время, — это системы с преобразованием на уровне синтаксической или семантико-синтаксической структуры. В последнее время именно этап преобразования привлекает наибольшее внимание специалистов. Сюда переместилось решение многих трудных проблем перевода.

По странам доклады по машинному переводу распределились следующим образом: Япония — пять докладов, Нидерланды — четыре, Канада, США, ФРГ, Англия, Испания — по два доклада, Франция, Бельгия, Дания, Израиль, КНР — по одному докладу.

Япония в настоящее время является безусловным мировым лидером в машинном переводе. По оценке специалистов, сейчас в этой области в Японии работает 800—900 человек, из них около 60% в фирмах, остальные — в государственных учреждениях. Число специалистов по машинному переводу во всех остальных странах мира также составляет 800—900 человек. Подавляющее большинство японских систем делается для перевода в паре языков: японский — английский (в ту или иную сторону или в обе стороны). В 1984 г. система ATLAS-I стала первой коммерческой японской системой. За последние три года на рынок вышли системы: PIVOT, Hicats/JE, Pensee, MELTRAN. Крупнейший правительственный проект (Mu-проект) имеет целью перевод аннотаций статей по науке и технике с английского на японский и с японского на английский. Основная исследовательская часть в нем делалась специалистами из университета Киото, в разработке участвовали Electrotechnical Laboratories (Tokyo), Japan Information Center for Science and Technology—JICST, Research Information Processing System under the *Agency of Engeneering Technology. В 1982—1986 гг.* была разработана система-прототип, в *1986—1990* гг. должна быть создана практически действующая система. Работа над этой второй очередью перешла в основном в JICST, а группа университета Киото переключалась на новое направление исследований. Это новое направление, развиваемое в основном в Японии, — перевод диалога в реальном времени. Диалог ведется либо через терминалы ЭВМ, либо даже устно по телефону. В последней постановке задача особенно трудна, так как ко всем трудностям перевода добавляются трудности анализа устной речи. По оценке специалистов, системы машинного перевода устной речи могут появиться через 15 лет.

В докладе ведущих специалистов из Киото **J. Tsujii, M. Nagao** «Перевод диалога vs. перевод текста — подход, основанный на интерпретации» рассматривались следующие особенности перевода диалога по сравнению с переводом текстов. (Имелся в виду диалог через терминалы, т. е. последовательность письменных сообщений.) Во-первых, пока ставится задача перевода диалога только по определенной узкой тематике, например, резервирование мест в гостинице, участие в конференции и т. п. Это облегчает понимание сообщения и позволяет разделить содержащуюся в сообщении информацию на релевантную для данного диалога и несущественную. Чтобы слушатель мог правильно понять собеседника, важно передать в переводе релевантную информацию, причем сохранение формы сообщения не обязательно. Иначе говоря, это скорее пересказ на другом языке, чем точный перевод. В докладе рассматривались способы выделения релевантной информации.

Во-вторых, особенность перевода диалога состоит в том, что его участники могут уточнять поступившие к ним сведения, задавая вопросы, причем переводческая система также может задавать вопросы в трудных для нее местах. Вместе с тем, важно, чтобы перевод шел достаточно быстро и не задерживал общение. Докладчики высказали утверждение, что система для перевода диалога не может создаваться как модификация систем перевода текстов, а должна разрабатываться с учетом указанной специфики.

Проблеме перевода устного диалога, ведущегося по телефону, был посвящен доклад японских специалистов **K. Kakigahara, T. Aizawa** (ATR Interpreting Telephony Research Laboratories, Osaka), названный «Пополнение японского предложения вставкой служебных слов на базе значащих слов». Система для перевода устного диалога объединяет систему распознавания речи, собственно систему перевода и систему синтеза речи. Известно, что системы распознавания речи работают с ограниченной точностью, поэтому для системы перевода в этой комбинации возникают дополнительные трудности. Особенно трудно распознаются короткие служебные слова, среди которых много сходных по звучанию; длинные значащие слова распознаются лучше. Поэтому докладчики предложили способ уточнения распознавания путем построения правильного предложения из поступивших значащих слов и сравнения его служебных слов с теми гипотезами, которые предлагает для них система распознавания. Тематика диалога установлена заранее: конференция (сроки, тематика, условия регистрации, резервирование гостиниц и т. п.). Работа находится в начальной стадии.

Еще одна система перевода японо-английского диалога (в реальном времени, с набором сообщений на терминалах ЭВМ и передачей через спутник между Японией и Швейцарией) была освещена в докладе **H. Nogami, Y. Yoshimura, S. Amano** (Research and Development Center of Toshiba Corporation) *«Анализ с забеганием в системе, переводящей в реальном времени».* Основное внимание докладчики уделили способу уменьшения числа возникающих при анализе гипотез *на основе забегания вперед: система* проверяет, нет ли спереди (в еще не проанализированной части поступившего сообщения) определенных опорных элементов, необходимых для принятия тех или иных гипотез об анализируемой части. Наличие или отсутствие их элементов может приводить к отбрасыванию как отдельных гипотез, так и целых групп. В приводимых примерах авторы используют то обстоятельство, что диалог обычно ведется достаточно короткими предложениями, в отличие от тех, которые встречаются в написанных текстах. Эксперимент связи между лабораторией фирмы Toshiba и Женевой был проведен в 1987 г.

Группа специалистов из университета Киото **J. Tsujii, Y. Muto, Y. Ikeda, M. Nagao** представила доклад на тему «Как получить предпочтительное прочтение при анализе естественных языков». В нем рассматривалась система анализа текстов, которая может, наряду с правилами запрещения, использовать правила предпочтения, причем она может в процессе анализа либо отбирать предпочтительные пути из числа альтернативных, либо построить все возможные структуры в порядке предпочтительности.

Специалисты из университета префектуры Осака **F. Nishida, S. Takamatsu, T. Tani, T. Doi** в докладе «Обратное воздействие информации об исправлениях при постредактировании на систему машинного перевода» рассказали о попытке автоматизировать учет исправлений, вносимых постредактором в текст, возникающий в результате работы переводческой системы. Авторы предложили для этого использовать систему, которая будет переводить в «обратном» направлении: если «прямая» система переводит с японского на английский, то «обратная» переводит отредактированный английский текст на японский. Промежуточные представления, возникающие при прямом и обратном переводе, сверяются, и на основе их сопоставления и учета

расхождений вносятся изменения в словарь и грамматику «прямой» системы. Работа находится в начальной стадии.

Серия докладов специалистов разных стран была посвящена совместному проекту стран ЕЭС — EUROTRA. Это наиболее амбициозный в мире проект по ожидаемому качеству результатов и наиболее крупный по числу работающих специалистов и по числу участвующих языков. В нем принимают участие около 100 специалистов Англии, Бельгии, Нидерландов, Греции, Дании, Ирландии, Италии, Люксембурга, ФРГ, Франции. Кроме того, имеется центральная группа, и секретариат в Женеве. Работа началась в 1978 г. Проект охватывает языки: английский, голландский, греческий, датский, итальянский, немецкий, французский. Решается вопрос о подключении испанского и португальского. Перевод предполагается с любого из названных языков на любой, что требует выработки единого промежуточного представления для этапа преобразования. Каждый язык будет иметь пять уровней представления: текст, морфологическое представление, синтаксическое представление в терминах компонент, синтаксическое представление в терминах синтаксических отношений, представление в терминах глубинных падежей. Основной формализм — недетерминированный преобразователь «дерево — дерево». Не предполагается использовать экстралингвистические знания или когнитивные модели типа тех, которые разрабатываются в исследованиях по искусственному интеллекту; не учитываются свойства дискурса. Грамматики для отдельных языков пока слабо развиты. Созданный малый прототип работает очень медленно. Видимо, к 1990 г. система-прототип готова не будет, так же как не будет к 1993 г. работающей системы (сроки, установленные при начале работы). Тем не менее, Комиссия Европейского парламента, обследовавшая работу, рекомендовала закончить в 1988 г. исследовательскую стадию, больше внимания уделить эффективной реализации и начать создавать практически действующую систему.

Проекту EUROTRA было посвящено пять докладов.

Доклад **P. Schmidt** (ФРГ) «Описание синтаксиса немецкого языка в формализме, предназначенном для машинного перевода» включал как описание используемого формализма, так и изложение принципов представления синтаксиса немецкого языка с его помощью.

В докладе **A. Bech, A. Nygaard** (Дания) «E-схема: формализм для переработки естественных языков» был описан формализм для преобразования от уровня к уровню в многоуровневой системе перевода. E-схема состоит из двух компонент: транслятора и генератора, каждый из которых реализует работу совокупности правил, заданных в декларативной форме. Транслятор преобразует дерево в объект, названный частичным описанием, содержащим информацию о признаках узлов в виде множества пар: атрибут — значение. Генератор дает на выходе дерево, т. е. устанавливает для узлов отношения порядка и подчинения. Такое разделение позволяет упростить правила и сократить их число.

В докладе **E. N. Steiner, J. Winter-Thielen** (ФРГ) «О семантике феномена фокуса в EUROTRA» рассматривалась проблема выделения фокуса в немецком предложении и связь расположения фокуса с порядком слов и семантической интерпретацией предложения.

F. Van Eynde (Бельгия) в докладе «Анализ времени и вида в EUROTRA» предложил формализм для описания временных отношений между компонентами дискурса и осветил способ приведения различных форм выражения времени и вида, встречающихся в языках, входящих в проект EUROTRA, к единой форме, используемой на этапе преобразования.

Близким к предыдущему по теме был доклад **M. Meya, J. Vidal** (Испания) «Интегрированная модель для обработки времени в системах машинного перевода», где, в основном на примере испанского языка, разбиралось, как сведения о временных отношениях событий извлекаются из рассмотрения формы глагола, наречий и именных групп со значением времени.

Несколько докладов было посвящено этапу преобразования (этапу Transfer) при переводе. Как известно, в первых системах перевода (так называемых «прямых») анализ входного текста делался с ориентацией на выходной язык, и в сущности весь процесс перевода состоял из очень сложного преобразования (или почти весь, если выделялись морфологические анализ и синтез). Другой крайний случай — это переход через нейтральное относительно языков представление (interlingua). В интерлинговых системах оказалось трудным сохранять — в общем для многих языков представлении — сведения о поверхностной организации текста, которые важны для правильного перевода. В настоящее время наиболее подходящими для перевода представляются системы с преобразованием (transfer systems), в которых переход от языка к языку идет на уровне синтаксической или семантико-синтаксической структуры. В многоязычных системах уровень преобразования обычно глубже, чем в двуязычных.

В докладе **J. A. Alonso** (Испания) «Модель управления преобразованием в машинно-переводческой системе METAL» рассматривался способ выделения и записи сведений, позволяющих определять порядок работы на этапе преобразования. Модель находится в стадии экспериментальной проверки. Она создавалась для немецко-испанской версии METAL.

METAL для англо-немецкого и немецко-английского перевода — это коммерческая система, сделанная для фирмы Сименс, которая финансировала соответствующие разработки, начиная с 1978 г. Из коммерческих систем — это наиболее сложная система с преобразованием. Предполагается ее стыковка с многоязычным терминологическим банком фирмы Сименс, TEAM.

I. Golan, S. Lappin, M. Riman (Израиль) в докладе «Активный двуязычный словарь для машинного перевода» представили свой способ записи сведений о соотношении лексических единиц двух языков в виде правил, записанных в словарных статьях двуязычного словаря. Эта работа выполнена в рамках проекта MENTOR, в котором участвуют несколько Европейских научных центров IBM.

В докладе **E. van Munster** (Нидерланды) «Обработка сферы действия отрицания в Rosetta» речь шла о системе Rosetta, предназначенной для англо-голландского перевода (в обе стороны), а также перевода с английского и голландского языков на испанский. Были предложены правила установления сферы действия отрицания, что очень важно ввиду того, что схема расположения основных компонент для английского и испанского: субъект—глагол—объект, а для голландского: субъект—объект—глагол. Рассмотрена работа этих правил на этапе преобразования.

«Интерактивный перевод: новый подход» — такова тема интересного доклада **R. Zajac** (Франция). Предлагается создать систему, которая будет помогать пользователю получить перевод с его языка на чужой в процессе написания документа. Обнаруживая неоднозначности во входном тексте, система будет запрашивать автора, какое именно толкование он имел в виду, предлагая ему на выбор перифразы различных возможных пониманий его сообщения. Такой способ не требует от автора никаких знаний ни о системе, ни о лингвистических правилах, а только понимания его собственных намерений. Система находится в стадии разработки; создается набор правил перифразирования.

M. McGee Wood, B. Chandler (Великобритания) в докладе «Машинный перевод для знающего один язык» рассказали об интерактивных системах NTRAN и AIDTRANS, которые предназначены для англоязычного

пользователя и задуманы с установкой: «перевод с японского и на японский без знания японского языка». Система AIDTRANS предназначена для японо-английского перевода, имеет словарь 6000 единиц, и дает перевод, близкий к подстрочнику. Анализ в основном сводится к снятию многозначности по линейному контексту и опирается на словарь, в который стараются поместить как можно больше разнообразной информации о специфических свойствах лексических единиц. Анализ дает несколько возможных результатов на выбор пользователю. Система-прототип написана на языке *C* (Си) и реализована на микроЭВМ. Другая система — NTRAN — предназначена для перевода с английского на японский. Она ориентирована на ограниченный входной язык. В процессе интерактивного предредактирования пользователь снимает неоднозначности входного текста. Со временем предполагается добавить выбор японских переводных эквивалентов на основе предлагаемых толкований.

A. Melby (США) в докладе «Лексическое преобразование: между скалами входа и выхода» рассказал о методике создания двуязычных баз данных с использованием конкордансов, предназначенных для улучшения выбора лексических единиц на этапе преобразования.

Доклад «Статистический подход к переводу языков» **P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jerlinek, R. Mercer, P. Roossin** (США) был посвящен использованию статистических методов извлечения полезной для перевода информации из больших двуязычных корпусов текстов (являющихся переводами друг друга) на английском и французском языках.

P. Isabelle, M. Dymetman, E. Macklovitch (Канада) в докладе «GRITTER — система для перевода отчетов сельскохозяйственного рынка» рассказали о системе перевода еженедельных отчетов Министерства сельского хозяйства с английского на французский и наоборот. GRITTER — это система типа transfer с преобразованием на уровне семантического представления, имеющего вид направленного ациклического графа (не обязательно дерева). Анализ и синтез предполагается сделать обратимыми. Разрабатываются синтаксический и семантический уровни, готова морфология для обоих языков, подготовлен словарь в 500 слов.

В докладе «Стилистическая грамматика для перевода» **C. DiMarco** и **G. Hirst** (Канада) рассматривались вопросы улучшения стиля при англо-французском и французско-английском переводе.

Z. Chen и **Q. Gao** (КНР) в докладе «Система англо-китайского машинного перевода IMT/EC» обрисовали строение системы и процесс перевода. Система с преобразованием, пока справляется с простыми случаями, рассчитана на постредактирование. Описанная в этом докладе система — одна из многих, разрабатываемых в настоящее время в КНР (в основном для перевода с английского на китайский), где после десятилетнего застоя начался новый подъем и в настоящее время начинается переход от экспериментальных систем к практически действующим. Так, система TRANSAR должна была выйти к концу 1988 г. на рынок, системы ISTIC-I и MT-IR-ER находятся в стадии проверки в экспериментах.

Три доклада голландских специалистов: «Машинный перевод: языковые сети (versus язык-посредник)» (**P. C. Rolf**); «DLT — индустриальный проект исследования и разработки многоязычного машинного перевода» (**T. Witkam**) и «Неявность, как ведущий принцип в машинном переводе» (**K. Shubert**) — были посвящены проблеме языка-посредника при многоязычном переводе, причем два последних доклада — системе многоязычного перевода, использующей эсперанто в качестве языка-посредника.

Доклад **J. L. Beaven, P. Whitelock** (Великобритания) «Машинный перевод, использующий изоморфные унифицированные категориальные грамматики (УКГ)» содержал описание формальной модели, пока не воплощенной в переводческую систему. Предлагается использовать УКГ в рамках изоморфных грамматик. Эти грамматики возникли как развитие идей Монтегю, в них выдерживается параллелизм между синтаксическими правилами входного и выходного языков.

Теме «Лексика» было посвящено восемь докладов. Рассматривалась организация сведений в словарях, используемых в вопросо-ответных системах, способы описания метафор. Три доклада были посвящены работам с большими словарями, которые существуют на машинных носителях, но создавались для человека, а не для автоматических систем,— это Longman Dictionary of contemporary English и Italian Machine Dictionary. Рассматривались различные подходы к их автоматизированному преобразованию в словари для систем автоматической обработки текста, а также способ автоматического построения тезаурусов на их базе.

По теме «Синтез» было прочитано семь докладов — это в основном работы по синтезу текстов, выполняемые для реализации человеко-машинного диалога на естественном языке с экспертными системами, базами данных и т. п. Один доклад был посвящен проблемам выбора лексики, в частности, выбора слова из ряда синонимичных (например, *boy, kid, child, youth*) на основе признаков слов и их соотношения с контекстом. В других — основное внимание уделялось вопросам построения структуры синтезируемого предложения (т. е. этапам семантического и синтаксического синтеза). Оригинальный подход к синтезу текстов был освещен в докладе специалистов из Саарбрюккена: в их системе XTRA, предназначенной для общения с экспертной системой на немецком языке, допускается не только словесное описание, но и прямое указание объектов на экране.

В докладе канадских специалистов сообщалось об экспериментальной реализации системы перифразирования, использующей как семантические, так и синтаксические правила, в соответствии с известной моделью «Смысл ↔ Текст» (см. ниже). Был также доклад, посвященный выбору стратегии при создании описания на основе определенной базы знаний, содержащей объекты и их отношения (на примере описаний интерьера квартиры). Стратегии в данном случае составляются из определенных наборов движений («после описания одного объекта переходить к тому, который расположен справа от первого» и т. п.).

По теме «Анализ и синтез речи» были заслушаны восемь докладов: о синтезе речи по графической записи для голландского языка; об опознавании границ слов в последовательности фонем на материале английского языка; об использовании модификации марковской модели при анализе речи, а также об использовании деревьев зависимостей и LF-грамматик при распознавании, сочетающемся с синтаксическим анализом.

Из пяти докладов, заслушанных на секции «Математическое обеспечение», три были посвящены программам, построенным для облегчения разработки и проверки грамматик, правила которых записаны в соответствующих формализмах,— так называемым Grammar developement environment (GRE). В одном из докладов рассматривался вопрос ускорения работы совокупности правил благодаря предварительному установлению определенных отношений между ними, учитывающих, в частности, что некоторые правила могут применяться только после определенных предшественников.

По теме «Синтаксис и морфология» шесть из десяти докладов были посвящены досинтаксическому уровню. В них рассматривались алгоритмы перехода от фонемной записи к морфемной, а также различные способы морфологического анализа для языков: финского, не-

мецкого, арабского и семитских с их специфическими трудностями из-за наличия интерфиксов.

Два доклада были посвящены системам обучения с помощью ЭВМ, хотя естественно было бы отнести эти доклады к соответствующей секции. В обоих случаях речь шла о системе, которая, получив на входе текст с ошибками (орфографическими, согласования и т. п.), выявляет их и предлагает способ исправления. Орфографические ошибки выявляются с помощью словаря и морфологического анализа, ошибки согласования — путем попытки синтаксического анализа рассматриваемого предложения. Одна из этих систем создана японскими специалистами и используется для обучения студентов английскому языку. Она состыкована с системой перевода на японский, которая после исправления ошибок дает перевод. Другая система создана специалистами из CNRS (Франция, Марсель), ориентирована на тексты из области геометрии и предназначена для формулирования теорем на французском языке, которые затем доказываются с помощью компьютера.

Два доклада были посвящены частным вопросам синтаксического анализа; один из них — проблемам снятия омонимии в голландских текстах, другой — проблемам анализа и восстановления эллипсиса в сочинительных конструкциях на материале русского языка (**И. А. Большаков**, СССР). В этом докладе была дана классификация всех обнаруженных «формул» русского сочинительного эллипсиса и предложены основы алгоритма его восстановления, опирающегося на грамматическое и семантическое сходство оставшихся частей.

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Компьютерная лингвистика — наука, об интенсивном развитии которой в нашей стране пока, к сожалению, говорить не приходится. Между тем, как сказал один из участников круглого стола на тему «Трудности автоматической обработки естественного языка», отношение компьютерной лингвистики к теоретической можно сравнить с отношением инженерного дела к математике. В этом смысле можно считать, что у нас почти нет «нженерного дела», а в «математике» развиваются лишь некоторые области. Поэтому интересно получить представление о том, какие области теории уже освоены «инженерным делом». Это тем более интересно, что «инженерное дело» продвинулось уже так далеко, что создаются системы (по типу экспертных), которые могут предоставить пользователю подробную информацию о «лингвистической мощности» любой интересующей его системы обработки языковой информации, т. е. о том, с какими лингвистическими проблемами та или иная система справляется, что она «умеет делать» (Ср. Artificial Intelligence Measurement System — AIMS).

Представим себя пользователями такой экспертной системы и познакомимся прежде всего с тем, в какой области лежат интересы компьютерной лингвистики. Оказывается, что это в основном семантические проблемы самого широкого спектра, которые обсуждались на конференции, например:

— интерпретация именной группы в интенсиональном контексте (**L. Lesmo, P. Terenziani** — Италия);

— интерпретация предложений, содержащих указательные местоимения (**J. Gundel, N. Hedberg, R. Zacharski** — США);

— проблема соотношения пресуппозиции и мнения (**D. Horton, G. Hirst** — Канада);

— метонимия, метафора и аномалия (**D. Fass** — США);

— семантика фокуса (**Steiner E. H., Winter-Thielen J.** — ФРГ);

— разрешение анафорических отношений (**J. G. Carboneli, R. D. Broun** — США);

— семантика различных синтаксических трансформаций (например, трансформации пассивизации и некоторых других) (**D. Jurafsky** — США).

По-видимому, наиболее сильный эффект от работ, выполненных в рамках «компьютерной лингвистики», заключается в том, что проблемы, казавшиеся сугубо отвлеченными и теоретическими, во-первых, приобретают реальное практическое значение и, во-вторых, могут быть решены с помощью машины. Работающие системы разрешают анафорические отношения, «понимают» типы метонимии и делают многое другое — из области, как мы привыкли считать, «чистой» науки. Но как инженерное дело может иногда продвинуть математику, например, поставив перед ней новые задачи, так и компьютерная лингвистика связана с лингвистикой теоретической. Авторам систем приходится в довольно жестких условиях давать теоретические решения — машина не терпит размытых формулировок. Так, в системе EUROTRA принимается, что фокус — это то, что не презумпция. Спорное определение работает, решая те задачи коммуникативной организации высказывания, которые интересуют авторов. А это для компьютерной лингвистики наиболее существенно, так как «если в лингвистической теории понятия «решить проблему», вообще говоря, нет, то в компьютерной лингвистике есть: это значит получить правильный результат в реальное время» (**K. Jensen**, исследовательский центр фирмы IBM). Если такой результат (в данном случае, высказывание с правильной коммуникативной организацией) не получен, то система работает плохо.

Значит ли это, что задачи лингвистики и компьютерной лингвистики все-таки несколько разные? Конечно, как афористично сформулировал J. Tsujii (Япония), теоретическую лингвистику в основном занимают правила и закономерности, по которым грамматическое отличается от неграмматического, тогда как в компьютерной лингвистике разрабатываются правила, по которым одни грамматические формы отличаются от других грамматических.*

Это, конечно, разные задачи, Нам, однако, интересны точки соприкосновения. Вот традиционная область лингвистики — лексикография. На конференции были представлены работающие системы, содержащие, например, такую лексикографическую информацию:

1) Информация при глаголе о том, к какой семантической группе он принадлежит — группе неконтролируемых событий (Events) или действий (Actions), — чтобы правильно устанавливать анафорические отношения в случаях типа *I did it* или *It happend*: «Anaphoric Reference to Evence and Actions: a Representation and its Advantages» (**Schuster E.** — США).

2) Информация о «скрытых рефлексивах» типа *dress* ('одевать'/'одеваться'), *wash* ('мыть'/'мыться') — ср. строго нерефлексивные *eat* 'есть', *read* 'читать' и др. Это позволяет диалоговой системе Start (**B. Katz, B. Levin** — (Artificial Intelligence Laboratory, Massachusetts Institute of Technology) по-разному реагировать в ситуациях *Sally ate the apple* и *David dressed the baby*: если на вопрос *Did Sally eat?* машина отвечает — *Yes*, то на вопрос: *Did David dress?* машина отвечает — *I don't know.*

* Между прочим, по мнению K. Jensen, компьютерную лингвистику должны (вопреки известному постулату, выдвинутому Хомским еще в 1957 г.), интересовать не только грамматические, не и не грамматические, формы: во-первых, потому, что то, что сегодня не грамматическое, завтра станет грамматическим (аргумент «диахронический»), а система должна быть рассчитана и на завтра, и, во-вторых, потому что реальный пользователь не всегда в ладах с грамматикой. Таким образом, нужны такие системы, которые умели бы интерпретировать и не вполне грамотные тексты.

3) Информация о реляционной структуре некоторых имен — система Lexical Conceptual Paradigm on the Semantic Interpretation of Nominals. Выделяются типы таких имен, например: результаты номинализации (*разрушение*), простые реляционные имена (*отец, мать*), артефакты (*книга, сигарета*), имена типа figure — ground (*порез, царапина* — каждое из них связано с именем соответствующей поверхности). Система понимает неоднозначность лексем типа *дверь* и *окно*, обозначающих не только предмет, но и проем, и, вводя реляционную структуру, успешно анализирует как предложения типа *Книга красного цвета*, так и типа *Книга займёт у тебя пять дней* (**J. Pusteyovsky, P. Anick**).

Интересно, что семантическую информацию о лексемах, необходимую машине, можно и извлекать с помощью машины. Впрочем, такая работа может быть сделана и вручную: ироничный J. Wilks (США), подсчитав, что на это уйдет два человекостолетия, предлагает свой проект в трех разных вариантах, где эта процедура полностью (или почти полностью — на 95%) компьютеризована.

Особо отметим японо-американскую работу «Understanding Stories for Animation» (**H. Shimazu, Y. Takashima, M. Tomono**), которая представляется интересной во многих отношениях. В частности, лингвиста-лексикографа она может побудить к исследованию таких явлений внутренней организации текста, которые до сих пор оставались совершенно не замеченными. Задача: по некоторому очень простому тексту (например, тексту детских сказок про зверей) порождать графическое изображение ситуации («компьютерная мультипликация»).

Рассмотрим следующий текст:

(1) *Заяц бежал от черепахи.*
(2) *Заяц посмотрел назад.*
(3) *Заяц сказал: «Черепаха меня ни за что не поймает».*
(4) *Заяц улегся на траву.*

Процедура порождения графического изображения показала, что человек как-то понимает неполные тексты. Для носителя языка такой текст совершенно однозначен. Ему не придет в голову представить себе, что заяц лег на бегу, причем с головой, повернутой назад. Между тем, чтобы такого не нарисовала машина, необходимо, по крайней мере, чтобы она поняла, что между (1) и (2) заяц остановился, а между (2) и (3) — повернул голову вперед.

Другой текст:

(1) *Черепаха бежала.*
(2) *Она бежала, и пыль стояла столбом.*
(3) *Она бежала, и пот лил с нее ручьем.*
(4) *Она взбежала на гору и остановилась.*

Надо понять, что *пыль столбом* — это следствие бега, и *пот ручьем* — тоже; что когда пот лил ручьем, пыль все еще поднималась столбом (хотя это и не сказано), и наоборот, что когда черепаха остановилась, то пыль исчезла и пот перестал литься.

Оживленную дискуссию вызвал доклад **P. Saint-Dizier** (Франция) «Немонотонная логика, естественный язык и обобщенные кванторы», посвященный лингвистическим применениям немонотонной логики (default logic), основные идеи которой были заложены в работах Дж. МакКарти* и которая в настоящее время широко используется в так называемом искусственном интеллекте, в частности, в теории представления знаний и при построении экспертных систем. Задача немонотонной логики состоит в формализации таких рассуждений, которые апеллируют к типичным свойствам объектов, т. е. к свойствам своего класса, а возможные исключения игнорируются, т. е. не принимаются во внимание до тех пор, пока не будет предъявлена информация о том, что в данном случае мы имеем дело именно с нетипичным, исключительным объектом (отсюда второе название для такого типа выводов — выводы по умолчанию, by default, т. е. при отсутствии информации об обратном). Немонотонная логика позволяет из утверждения x — *птица* сделать заключение о том, что x *летает*, и придерживаться этого заключения в ходе рассуждения до тех пор, пока не будет получена, например, информация о том, что x — *страус*. Правила вывода в немонотонной логике имеет следующий вид: «Если $P(x)$ и нет прямых указаний о том, что $P(x)$ ложно, то утверждение $P(x)$ может быть принято как истинное, хотя и с неполной степенью уверенности».

Автор предлагает ряд правил, позволяющих уточнить семантику кванторных слов (таких как *обычно, в типичном случае, большинство, подавляющее большинство, все, почти все, много, мало, несколько, немного* и др.), а также значение контекстно зависимых детерминативов, превосходной степени прилагательного, прилагательных типа *превосходный, совершенный, идеальный,* а также безличных глагольных конструкций. Некоторые правила вывода в предлагаемой естественной логике вызвали у присутствующих сомнение — в частности, правило, по которому из посылок

Все млекопитающие являются животными и
Большинство животных — травоядны

можно прийти к заключению

Большинство млекопитающих — травоядны.

Во втором докладе о немонотонных логиках (авторы **U. Zernik, A. Brown**) (США), который назывался «Немонотонные рассуждения в языковых процессорах», говорилось о применении методов немонотонной логики при синтаксическом анализе. Речь шла об анализе предложений с локальной неоднозначностью, которая может быть обнаружена лишь на последней стадии анализа: это так называемая garden path phenomena — явление, которое может быть продемонстрировано на примере *The boat floated by the river sank*; структура этого предложения такова, что оно «вводит в заблуждение», поскольку при прочтении предложения *floated* неправильно идентифицируется как сказуемое при *boat* (хотя оно может быть и причастием-определением), и ошибка обнаруживается лишь тогда, когда читатель доходит до последнего слова во фразе, *sank*, которое представляет собой однозначное сказуемое, но остается не присоединенным ни к какому подлежащему. Предложения этого типа предлагается трактовать по аналогии со страусами. До определенного момента эти предложения анализируются ошибочно, но по таким правилам, которые позволяют существенно быстрее находить правильную структуру для более типичных предложений.

Доклад «Референциальные свойства родовых термов» (**Е. В. Падучева,** СССР) на секции «Семантика» был посвящен родовым именам — выражениям, представляющим давнюю загадку для лингвистической теории референции. Предложено деление родовых термов на две группы: в одних контекстах родовой терм обозначает соответствующий класс (например, *кит — млекопитающее; ягуары в Южной Америке вымирают*), а в других (*норвежцы — высокого роста; соната обычно состоит из четырех частей*) является референциально неполной именной группой — общим именем с адвербиальной квантификацией, имплицитной или эксплицитной.

Один из докладов был посвящен системе, выполненной на базе модели «Смысл⇔Текст», известной в нашей стране по работам И. А. Мельчука, А. К. Жолковского, Ю. Д. Апресяна и др. Это система GOSSIP (Generation of Operating System Summaries In Prolog — **L. Iordanskaya, R. Kittredge, A. Palguère**), которая

* См. также: Reiter R. A Logic for default reasoning. — Artificial intelligence. V. 13, p. 81—132, 1980.

использует многоуровневую модель с правилами перехода от смысла к тексту для автоматического порождения сообщений о работе операционных систем. Порождение текста предлагает некоторый «долингвистический» уровень представления информации. Отличие данной системы от других систем того же типа заключается в том, что этот «долингвистический» уровень — уровень планирования будущего текста и выбора его содержания — не совпадает с семантическим уровнем. Первый ориентирован на заданную предметную область (в данном случае речь идет о протоколах операционных систем). Второй — семантический уровень — в конечном счете определяется лексикой и грамматикой конкретного языка. «Долингвистическое» представление называется концептуально-коммуникативным (Conceptual Communicative Representation), так как система предполагает в качестве важнейшего компонента коммуникативную структуру, с выделением темы и ремы для каждого сообщения (под сообщением — message — понимается полный ответ на вопрос). Тема соответствует объекту, находящемуся в фокусе внимания говорящего; рема соответствует значению определенного параметра этого объекта. Таким образом, рема — это то, что сообщается о теме.

Коммуникативное представление — это новый элемент в модели «Смысл⇔Текст», которая прежде ориентировалась только на задачу машинного перевода и в качестве глубинного имела только семантический уровень представления. В системе GOSSIP теоретические разработки, касающиеся коммуникативной структуры текста, получают, так сказать, свое «внедрение», а именно: заданная коммуникативная структура сообщения позволяет в дальнейшем правильно выбрать вершину глубинно-синтаксического дерева, правильно выбрать лексему предиката в данном сообщении (так, коммуникативная структура делает в каких-то случаях однозначным выбор между лексемами типа *покупать* и *продавать*), правильно выбрать поверхностно-синтаксическое оформление предложения (активную или пассивную конструкцию и т. п.).

В настоящее время система находится в экспериментальной стадии разработки: она располагает словарем в несколько десятков слов и не очень большим синтаксисом, однако и словарь, и синтаксис постоянно расширяются, с тем чтобы полнее использовать возможности модели «Смысл⇔Текст» — в первую очередь правила глубинно-синтаксического перефразирования с использованием лексических функций.

Авторы доклада, делая краткий обзор возможностей модели «Смысл⇔Текст», отмечают, что эта модель отличается от других тем, что в ней «центр тяжести приходится на словарь».* И дело здесь не только в изобретении абсолютно нового аппарата лексических функций, а в том, что модель потребовала научной организации словарной статьи, всей лексикографической информации, и, с другой стороны, прочно связала словарь с тем, что принято называть «грамматикой» — системным описанием языка в целом. Понятие «интегрального описания» языка (словарь+грамматика во взаимодействии) стало для московской школы лингвистической семантики привычным. Однако, как показало обсуждение на круглом столе «Взаимодействие словаря и грамматики в машинном переводе», связь этих двух компонентов языкового описания вовсе не является очевидной истиной для мировой вычислительной лингвистики.

Прошедший COLING-88 дал возможность познакомиться с разными направлениями исследований современной компьютерной лингвистики. Заслуживает благодарности безупречная организация конференции, предоставлявшая равные возможности и для выступлений, и для дискуссий, и для других мероприятий.

Образцовым следует признать и издание трудов конференции, составивших два тома докладов с подробным справочным аппаратом.** Все это стало возможным прежде всего благодаря усилиям Национального Венгерского оргкомитета — председатель B. Dömölki, секретарь D. Vargha, члены: T. Gergely, F. Kiefer, F. Papp и другие.

---

* Эта лексикографическая направленность получила еще более глубокое развитие в дальнейших разработках модели в группе Ю. Д. Апресяна (Институт проблем передачи информации АН СССР).

** Preceedings of the 12-th International Conference of Computational Linguistics (COLING-88), Budapest, 1988, V. 1—2.

Материал поступил в редакцию 26.12.88.

Редактор Т. Н. Лаппалайнен

Технический редактор Л. И. Хоченкова

Сдано в набор 24.02.89 Подписано в печать 14.03.89 Т—02711

Формат бумаги 84×108¹/₁₆ Бум. тип. № 2 Литературная гарнитура Высокая печать

Усл. печ. л. 4,20 Усл. кр.-отт. 4,86 Уч.-изд. л. 5,78 Тираж 6597 экз. Заказ 1450 Цена 30 коп.

Адрес редакции: 125219, Москва, А-219, Балтийская ул., 14. Тел. 152-66-71

Производственно-издательский комбинат ВИНИТИ,
140010, Люберцы, 10, Московской обл., Октябрьский пр., 403