УДК 81:322.4

ПРОБЛЕМЫ ПОСТРОЕНИЯ СИСТЕМ МАШИННОГО ПЕРЕВОДА

С.В. Фролов, Д.А. Паньков

ГОУ ВПО «Тамбовский государственный технический университет»; ОАО «Тамбовская областная сбытовая компания», г. Тамбов

Рецензент В.Е. Подольский

Ключевые слова и фразы: информационная система; машинный перевод.

Аннотация: Рассмотрены проблемы создания систем машинного перевода с 1940 года. Описывается стандартная структура и проблемы разработки современных систем машинного перевода.

Перевод (по определению) – это деятельность, заключающаяся в передаче содержания текста на одном языке средствами другого языка, а также результат такой деятельности. Особое место в теории перевода занимает машинный перевод (**МП**) – научная и одновременно технологическая дисциплина, связанная с наукой о переводе, а так же с компьютерной лингвистикой. Машинный перевод – это выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

При переводе текста с одного языка на другой человек опирается на свои знания о мире, использует самые разнообразные экстралингвистические знания (о физической природе мира, об обществе и его культуре и т.п.). Если же таким переводом занимается машина, то значительная часть экстралингвистической информации у нее просто отсутствует. И это приводит к тому, что до сих пор, после более чем 60-ти лет разработки систем машинного перевода, удовлетворительные результаты можно получить только при построении системы, рассчитанной исключительно на какую-то строго определенную область знания. Причем, участие человека-оператора для систем МП является обязательным, кроме некоторых случаев, например, когда все варианты входного текста можно описать заранее.

Фролов С.В. – доктор технических наук, профессор, заведующий кафедрой “Биомедицинская техника» ТамбГТУ; Паньков Д.А. – инженер 2-ой категории ОАО «Тамбовская областная сбытовая компания», г. Тамбов.

Развитие систем машинного перевода в том виде, в котором мы наблюдаем их сейчас, фактически началось в конце 40-х годов прошлого века. Огромную роль сыграло публикация меморандума «Перевод» Уореном Уивером (Warren Weaver) – директором отделения естественных наук Рокфеллеровского фонда (Rockefeller Foundation). Он впервые сформулировал принципы машинного перевода. Первая успешная реализация системы машинного перевода связана с «Джорджтаунским экспериментом», осуществленным на машине IBM-701 в 1954г. Программное и лингвистическое обеспечение этой попытки было весьма скромным: словарь состоял из 250 слов, а грамматика – из 6 синтаксических правил. Интересно, что перевод делался с русского языка на английский, и, возможно, этим объясняется последующая резкая активизация разработок в данной области в СССР. Уже к концу 1955г. в Институте научной информации Отделения прикладной математики МИАН СССР и некоторых других академических центрах были созданы и прошли тестирование программы МП на существовавшей тогда вычислительной базе (БЭСМ и "Стрела").

Следует заметить, что большинство исследователей истории МП не обращают внимание на тот факт, что параллельно с исследованиями в данной области происходило довольно бурное развитие как аппаратных средств – компьютеров, так и языков программирования. Если же рассматривать развитие систем МП «в комплексе», то становится ясно, что во многом степень их совершенства определялась именно уровнем развития аппаратных и программных средств.

Так, вплоть до первого угасания интереса к системам МП в 1959 году «узким местом» были вопросы хранения и доступа к данным в памяти, а также оптимизации алгоритмов из-за низкой производительности ЭВМ. В сравнении с современными персональными компьютерами, мэйнфреймы того времени обладали крайне низкой производительностью. Например, IBM 701, являвшаяся самой совершенной из ЭВМ того времени имела скорость вычислений до 17 тыс. операций в секунду. Для сравнения, не самый новый процессор для персонального компьютера Intel Core 2 Duo, выпущенный в 2006 году, обеспечивает производительность порядка 1,3 миллиарда операций в секунду.

Вплоть до конца 1970-х годов происходил «набор сил» для следующего рывка в развитии систем МП – разрабатывались лингвистические теории, совершенствовалась аппаратная часть и языки программирования.

Если в 1950-х годах лучшие мэйнфреймы имели объемы памяти, измерявшиеся килобайтами, то уже к середине 1980-х годов стандартный персональный компьютер имел 128 кб оперативной памяти с (возможностью расширения до 640 кб) при значительном уменьшении времени доступа к ней. Соответственно, к этому времени узким местом систем МП становится эффективность реализации алгоритма.

В это же время происходят значительные события и на фронте языков программирования. Так, появившийся в 1971 году язык программирования Си очень быстро становится популярным. К середине 1980-х концепция объектно-ориентированного программирования вместе с соответствующей

модификацией языка – C++, созданной Бьерном Страуструпом, в значительной степени меняет принципы проектирования программного обеспечения.

Такое бурное развитие, как аппаратных, так и программных средств не могло не сказаться на развитии систем МП. К началу 1990-х годов появляются первые системы МП промышленного уровня как в России, так и за рубежом.

Уже к началу 2000-х годов прогресс в области аппаратного обеспечения и языков программирования приводит к возможности практической реализации многих наработок лингвистов. Существующие процессорные мощности позволяют реализовать алгоритмы практически любой сложности. Кроме того, появляются новые методы построения больших программных комплексов.

Однако все это никак не отражается на отрасли МП. Большинство существующих систем по-прежнему используют алгоритмы, разработанные еще в начале 1990-х годов (а некоторые системы – и более ранние).

Основной проблемой при разработке новых алгоритмов является проверка их эффективности. И вот тут-то и возникают основные сложности.

В большинстве систем МП обработка переводимого текста, так или иначе, проходит следующие этапы анализа: лексический (графематический анализ, токенизация), морфологический, фрагментационный, синтаксический и семантический.

В некоторых системах часть этапов может отсутствовать или же не выделяться в явном виде.

Любая разработка обычно проходит две стадии. Первая стадия – это проверка работы программы на подготовленных особым образом исходных данных. На второй стадии осуществляется проверка на реальных данных, коррекция которых предварительно не производится.

Проблема в том, что для проверки работы, например, алгоритма синтаксического анализа на реальных данных необходимо реализовать предыдущие три этапа (это если не учитывать то, что необходимо разработать и интерфейс для ввода данных). А современные алгоритмы семантического анализа, как правило, работают с большим количеством статистической информации, которую тоже надо собрать, и на это тоже требуется время.

В Интернете практически не существует исходных кодов программ, осуществляющих перечисленные этапы анализа, а те, что есть – обычно обладают существенными недостатками. С другой стороны – есть огромное количество публикаций по каждому из этапов анализа.

Возникает ситуация, когда для проведения эксперимента необходимо львиную долю времени потратить на перевод в программный код словесных описаний алгоритмов, реализация которых мало сказывается на конечной задаче, однако решение ее без этих алгоритмов в принципе невозможно.

Возможным решением является приобретение готовых библиотек, реализующих требуемую функциональность. Однако цена таких библио-

тек велика и не всякий коллектив разработчиков (особенно в России) может себе это позволить.

Кроме того, на отдельные алгоритмы существуют еще и лицензионные ограничения. Это означает, что в случае использования данных алгоритмов в разрабатываемой коммерческой системе МП с каждой проданной копии придется осуществлять лицензионные отчисления.

В целом, на разработку интерфейса пользователя, реализацию и отладку первых трех этапов анализа может потребоваться год и более. Причем, время это будет фактически потрачено впустую.

Современный подход к созданию крупных программных комплексов позволяет решить эту проблему быстрым и дешевым способом.

Необходима разработка модульных систем МП, где каждый модуль будет осуществлять один из этапов обработки текста. Такой подход позволит совершенствовать алгоритмы анализа без разработки системы МП с нуля и сосредоточить все усилия разработчиков именно в выбранном направлении исследования.

На рис. 1 показана разработанная авторами статьи модульная система МП.

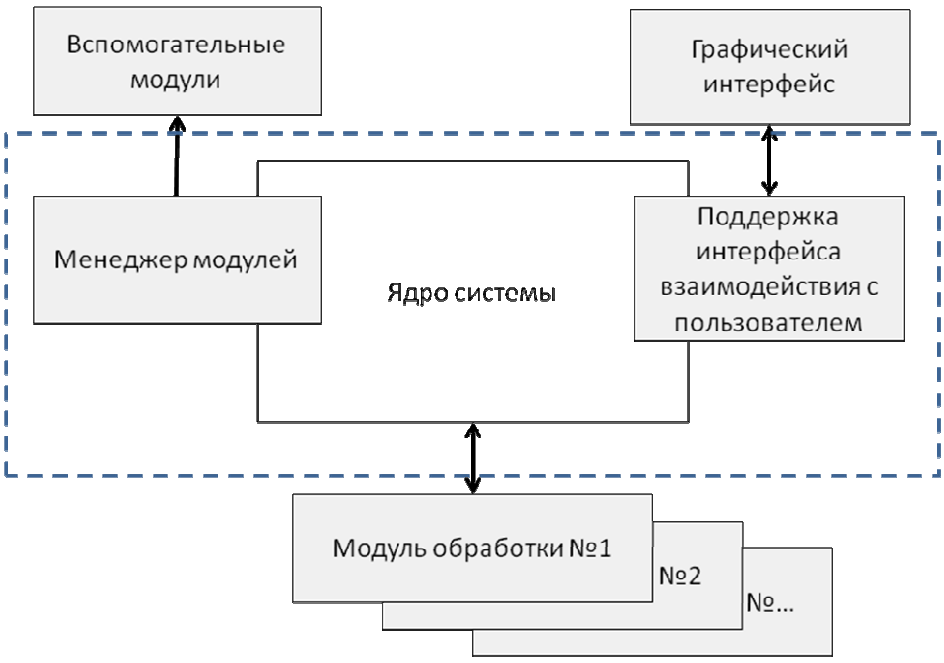

Рис. 1. Схема модульной системы МП

В процессе работы алгоритма перевода полученная от пользователя информация последовательно проходит через все модули обработки. Процесс обмена информацией между отдельными модулями и ее формат стандартизирован. Это позволяет легко подключать и отключать отдель-

ные модули, а также экспериментировать над различными реализациями алгоритмов анализа, выбирая из них лучшие.

Problems of Designing Systems for Machine Translation

S.V. Frolov, D.A. Panykov

Tambov State Technical University, Tambov;
"Tambov Regional Marketing Company" plc, Tambov

Key words and phrases: information system; machine translation.

Abstract: Problems of creating systems for machine translation beginning from 1940 are considered. Standard structure and problems of designing modern systems for machine translation are described.
