УДК 81'322; 81'322.4

МАШИННЫЙ ПЕРЕВОД В INTERNET

Каменева Н.А.
Московский государственный университет экономики, статистики и информатики

Исследованы теоретические и практические вопросы использования систем машинного перевода в современном едином многонациональном информационном пространстве. Отмечено, что наиболее часто машинный перевод в Интернет используется в поисковых системах для привлечения наибольшего количества пользователей к сайтам. Приведены наиболее распространенные системы машинного перевода PROMT Internet Translation, WebView, Lingvo и «МультиЛекс» и др. Автором предложены основным требованиям к успешному использованию программ машинного перевода: оперативность; гибкость; скорость; точность. Отмечено, что машинный перевод – уникальный гуманитарный инструмент, позволяющий преодолевать многочисленные проблемы общения и языковые барьеры.

Ключевые слова: многоязычная среда, браузер, Web – страница, система оптического распознавания, интерактивный режим, электронный словарь, поисковый сайт.

Постановка проблемы. Сейчас наблюдается новый всплеск интереса к системам машинного перевода (МП) в связи с развитием сети Internet. Миллионы людей, говорящих на разных языках, оказались в едином информационном пространстве. Доминирует в Сети английский язык, но есть пользователи, которые им не владеют, как, впрочем, есть множество Webстраниц, написанных не по-английски. Для облегчения просмотра страниц Internet на незнакомом пользователю языке появились дополнения к браузерам, которые осуществляют немедленный перевод выбранных пользователем фрагментов просматриваемой Webстраницы [4, с. 57]. Достаточно лишь выделить часть текста мышкой и перенести ее на специальную панель либо нажать указателем на специальную кнопку меню. Примером такого переводчика является система Web Trans Site фирмы ПРОМТ, созданная на базе программы Stylus, которая подключается как к браузеру Netscape Navigator, так и к браузеру Microsoft Internet Explorer.

Анализ последних исследований и достижений. В жизни современного общества важную роль играют автоматизированные информационные технологии. С течением времени их значение непрерывно возрастает. Но развитие информационных технологий происходит весьма неравномерно: если современный уровень вычислительной техники и средств связи поражает воображение, то в области смысловой обработки информации успехи значительно скромнее. Эти успехи зависят, прежде всего, от достижений в изучении процессов человеческого мышления, процессов речевого общения между людьми и от умения моделировать эти процессы с помощью компьютеров. Когда речь идет о создании перспективных информационных технологий, то проблемы автоматической обработки текстовой информации, представленной на естественных языках, выступают на передний план. Это определяется тем, что мышление человека тесно связано с его языком. Более того, естественный язык является инструментом мышления. Он является также универсальным средством общения между людьми – средством восприятия, накопления, хранения, обработки и передачи информации. Проблемами использования естественного языка в системах автоматической обработки информации занимается наука компьютерная лингвистика. Эта наука возникла сравнительно недавно – на рубеже пятидесятых и шестидесятых годов прошлого столетия. За прошедшие полвека в области компьютерной лингвистики были получены значительные научные и практические результаты: были созданы системы машинного перевода текстов с одних естественных языков на другие, системы автоматизированного поиска информации в текстах, системы автоматического анализа и синтеза устной речи и многие другие.

Компьютерная лингвистика – это область знаний, связанная с решением задач автоматической обработки информации, представленной на естественном языке. Центральными научными проблемами компьютерной лингвистики являются проблема моделирования процесса понимания смысла текстов (перехода от текста к формализованному представлению его смысла) и проблема синтеза речи (перехода от формализованного представления смысла к текстам на естественном языке). Эти проблемы возникают при решении ряда прикладных задач и, в частности, задач автоматического обнаружения и исправления ошибок при вводе текстов в ЭВМ, автоматического анализа и синтеза устной речи, автоматического перевода текстов с одних языков на другие, общения с ЭВМ на естественном языке, автоматической классификации и индексирования текстовых документов, их автоматического реферирования, поиска документов в полнотекстовых базах данных.

Лингвистические средства, создаваемые и применяемые в компьютерной лингвистике, можно условно разделить на две части: декларативную и процедурную. К декларативной части относятся словари единиц языка и речи, тексты и различного рода грамматические таблицы, к процедурной части – средства манипулирования единицами языка и речи, текстами и грамматическими таблицами. Компьютерный интерфейс относится к процедурной части компьютерной лингвистики.

Успех в решении прикладных задач компьютерной лингвистики зависит, прежде всего, от полноты и точности представления в памяти ЭВМ декларативных средств и от качества процедурных средств. На сегодняшний день необходимый уровень решения этих задач пока еще не достигнут, хотя работы в области компьютерной лингвистики ведутся во всех развитых странах мира (Россия, США, Англия, Франция, Германия, Япония и др.).

Тем не менее, можно отметить серьезные научные и практические достижения в области компьютерной лингвистики. Так в ряде стран (Россия, США, Япония, и др.) построены экспериментальные и промышленные системы машинного перевода текстов с одних языков на другие, построен ряд экспериментальных систем общения с ЭВМ на естественном языке, ведутся работы по созданию терминологических банков данных, тезаурусов, двуязычных и многоязычных машинных словарей (Россия, США, Германия, Франция и др.), строятся системы автоматического анализа и синтеза устной

речи (Россия, США, Япония и др.), ведутся исследования в области построения моделей естественных языков.

Важной методологической проблемой прикладной компьютерной лингвистики является правильная оценка необходимого соотношения между декларативной и процедурной компонентами систем автоматической обработки текстовой информации. Чему отдать предпочтение: мощным вычислительным процедурам, опирающимся на относительно небольшие словарные системы с богатой грамматической и семантической информацией, или мощной декларативной компоненте при относительно простых компьютерных интерфейсах? Большинство ученых считают что, второй путь предпочтительнее. Он быстрее приведет к достижению практических целей, так как при этом меньше встретится тупиков и трудно преодолимых препятствий и здесь можно будет в более широких масштабах использовать ЭВМ для автоматизации исследований и разработок.

Необходимость мобилизации усилий, прежде всего, на развитии декларативной компоненты систем автоматической обработки текстовой информации подтверждается полувековым опытом развития компьютерной лингвистики. Ведь здесь, несмотря на бесспорные успехи этой науки, увлечение алгоритмическими процедурами не принесло ожидаемого успеха. Наступило даже некоторое разочарование в возможностях процедурных средств.

В свете вышеизложенного, представляется перспективным такой путь развития компьютерной лингвистики, когда основные усилия будут направлены на создание мощных словарей единиц языка и речи, изучение их семантико-синтаксической структуры и на создание базовых процедур морфологического, семантико-синтаксического и концептуального анализа и синтеза текстов. Это позволит в дальнейшем решать широкий спектр прикладных задач.

Перед компьютерной лингвистикой стоят, прежде всего, задачи лингвистического обеспечения процессов сбора, накопления, обработки и поиска информации. Наиболее важными из них являются:

1. Автоматизация составления и лингвистической обработки машинных словарей;
2. Автоматизация процессов обнаружения и исправления ошибок при вводе текстов в ЭВМ;
3. Автоматическое индексирование документов и информационных запросов;
4. Автоматическая классификация и реферирование документов;
5. Лингвистическое обеспечение процессов поиска информации в одноязычных и многоязычных базах данных;
6. Машинный перевод текстов с одних естественных языков на другие;
7. Построение лингвистических процессоров, обеспечивающих общение пользователей с автоматизированными интеллектуальными информационными системами (в частности, с экспертными системами) на естественном языке, или на языке, близком к естественному;
8. Извлечение фактографической информации из неформализованных текстов.

Подробно остановимся на проблемах, наиболее относящихся к теме исследования.

Онлайновый перевод информации в Internet становится все более популярным. Internet стремительно превращается из преимущественно англоязычной в многоязычную среду, что вынуждает владельцев Web-сайтов предоставлять информацию на нескольких языках. Наиболее часто к услугам МП прибегают информационные и поисковые сайты, которые стремятся привлечь на свои страницы разноязычных пользователей. Так, на канадском информационно-поисковом портале infiniT (http://www.infiniT.com; http://fr.canoe.ca/) открылся новый сервис переводов. На сайте теперь доступен онлайновый перевод текста с английского и немецкого языков на французский язык и обратно.

Увеличение числа посетителей портала обусловлено возможностью онлайнового перевода Web-страниц. Для этого пользователю достаточно указать только адрес Web-страницы, выбрать направление перевода и нажать кнопку перевода. В результате через несколько секунд пользователь получает полностью переведенную Web-страницу с сохранением форматирования.

Новый сервис позволяет ликвидировать языковую проблему в канадском Internet, где в силу исторических особенностей широко используются два языка: английский и французский. Кроме того, онлайновый переводчик открывает доступ к сайтам на немецком языке тем жителям Канады, которые не владеют иностранными языками.

Сервис работает на базе серверного Интернет-решения компании ПРОМТ под названием PROMT Internet Translation Server version 2.0. Проект был реализован совместно с компанией Softissimo, которая занимается продвижением продуктов компании ПРОМТ под торговой маркой REVERSO.

Изложение основного материала. Интересной особенностью Web-сайтов, знакомящих с программами МП, электронными словарями и другими программами лингвистической поддержки, является то, что с работой многих программных продуктов можно познакомиться в интерактивном режиме, используя версию, установленную на сервере и имеющую шлюз для удаленного общения через Web-интерфейс. На сервере Web-издательства «ИнфоАрт» организована интерактивная демонстрация словарей Lingvo и «МультиЛекс». Вы можете ввести слово или словосочетание и мгновенно получить перевод, толкование, примеры употребления и устойчивые словосочетания.

При использовании универсальной переводящей программы WebTranSite или браузера WebView больше, чем других частей пакета PROMT Internet, то при этом возможно сэкономить немного денег, и можно приобрести все эти продукты по отдельности. В таком случае WebTranSite находит широкое применение, поскольку переводит небольшие фрагменты текста не только из Интернет, но и из офисных, почтовых и других программ, а также из системы интерактивной справки. Кроме того, WebTranSite подходит не только для перевода Web-страниц. Программа достаточно универсальна и позволяет обрабатывать фрагменты текста из любых приложений, в том числе из текстовых редакторов, электронных таблиц, органайзеров, браузеров. Программа также переводит тексты с английского, немецкого или французского языка на русский и обратно.

Что касается WebView, то это – инструмент искушенного «интернетчика», много времени проводящего на иноязычных серверах и не желающего пропустить ни одной крупицы информации. Программа WebView – это полноценный браузер для загрузки и просмотра Web-страниц. Используя его при работе с Интернет-документами на иностранном языке, можно даже отказаться от привычных Internet Explorer и Netscape Navigator. WebView – это своеобразный «гибрид» браузера Internet Explorer и системы машинного перевода

PROMT. Полученная в результате такого «скрещивания» программа позволяет переводить Web-страницы, полностью сохраняя их внешний вид. WebView может работать с тремя иностранными языками: английским, французским и немецким, причем переводит как с любого из них на русский, так и в обратном направлении. Функции перевода в этой программе шире, чем в WebTranSite, и их набор практически такой же, как в PROMT. Так, в дополнение к рассмотренным выше операциям со словарями WebView позволяет вводить в них новые слова и фразы [3, с. 27]. Чтобы облегчить вашу задачу, программа составляет список незнакомых ей выражений, встретившихся в тексте, подсчитывает количество появлений каждого из них и выводит результаты в специальном диалоговом окне. Выбрав любое из этих слов, можно либо запретить его перевод, либо добавить в пользовательский словарь. Единственный недостаток, который портит хорошее впечатление о WeView – это то, что программа иногда зависает при несовпадении направления перевода и языка оригинальной страницы.

Socrat Internet – это аналог «переводящего браузера» WebView. С ее помощью можно выполнять «синхронный» перевод Web-страниц с сохранением их форматирования. Однако если WebView по возможностям настройки опций перевода ничем не уступает профессиональной системе PROMT, то в Socrat Internet никаких средств управления этими функциями нет вообще. Браузер от компании «Арсеналъ» не позволяет подключать тематические словари, что сильно ухудшает качество перевода специальных текстов (их, кстати, в Интернет очень много). Простота использования, которой так добивались разработчики, отнюдь не пошла на пользу программе, потерявшей гибкость и управляемость, такие важные для систем машинного перевода. В итоге Socrat Internet существенно уступает продуктам «ПРОМТ» по многим параметрам, в том числе и по самому важному – качеству выходного текста.

Заключение. Стремительные потоки информационного обмена между вы-сокоразвитыми промышленными странами, лавина научно-технической документации, поступающая от производителей товаров и современных технологий, требуют совершенно нового подхода к проблеме перевода технической литературы. Выход один: максимально автоматизировать процесс, оставив человеку его творческую редакционную часть. В этом помогает система машинного перевода. Ее параметры должны удовлетворять четырем основным требованиям:

- оперативность;
- гибкость;
- скорость;
- точность.

Оперативность машинных систем – это возможность постоянного пополнения словарного запаса и создания новых тематических словарей. В этом параметре они значительно опережают привычные типографские издания различных словарей.

Гибкость – это возможность *«грубой настройки»* на конкретную предметную область (для этой цели служат специализированные словари) и *«тонкой настройки»* на конкретный текст, книгу или группу документов (модифицируемые пользовательские словари).

Скорость – возможность автоматического ввода и обработки текстовой информации с бумажных носителей. Только одна система оптического ввода текстов *(OCR-System – optical character recognition system)* ежедневно заменяет более десяти классных машинисток.

Точность – стилистически и грамматически правильная адекватная передача смысла исходного текста на язык перевода. Это наиболее *«уязвимое»* место систем машинного перевода. Однако столь явное улучшение качества перевода в поздних версиях систем машинного перевода, как например, PROMT, вселяет уверенность, что вскоре компьютер полностью примет на себя всю рутинную часть перевода.

Машинный перевод – это эффективное средство для просмотра и поиска информации на иностранном языке, и именно эта функция является главной при работе в Internet. Далее, в результате настройки на предметную область и интеграции с другими программами обработки документов средство машинного перевода позволяет автоматизировать получение перевода [1, с. 234]. И наконец, – это уникальный гуманитарный инструмент, позволяющий преодолевать проблемы общения в системах, работающих на разных языках. И пожалуй, самый главный, вывод состоит в том, что многие разработчики осознали: при создании программы машинного перевода кроме хорошо реализованной лингвистики необходима достойная программная реализация.

Список литературы:

1. Бархударов Л. С. Язык и перевод. – М.: Международные отношения, 1975. – 318 с.
2. Система перевода текста Magic Goody для Windows. Руководство пользователя. – С.-Петербург, фирма «ПРОМТ», 2010.
3. Система перевода текста WebTranSite. Руководство пользователя. – С.-Петербург, фирма «ПРОМТ», 2008. – 156 с.
4. Kameneva N.A. FOUNDATIONS OF COMPUTATIONAL LINGUISTICS. – Tutorial / Москва, 2014. – 110 с.
5. http://www/wikipedia.com
6. http://www.promt.ru
7. http://www.socrat.ru
8. http://www.translate.ru

Kameneva N.A.
Moscow State University of Economics, Statistics and Informatics

MACHINE TRANSLATION IN THE INTERNET

Summary
Theoretical and practical aspects of using machine translation systems in modern multinational information space are investigated. It is marked, that the most common machine translation in Internet is practiced in search engines to attract the largest number of users to their sites. The most common machine translation system PROMT Internet Translation, WebView, Lingvo and «MultiLex» and others are given and described. The author offers the basic requirements for successful use of machine translation programs: efficiency, flexibility, speed, accuracy. It is also noted that machine translation is a unique humanitarian tool to overcome many problems of communication and language barriers.

Keywords: multilingual environment, browser, Web – page, optical character recognition, interactive regime, electronic dictionary, a search site.