УДК 62-506.1

СИСТЕМА ПОИСКА, АНАЛИЗА И ОБРАБОТКИ МУЛЬТИЛИНГВИСТИЧЕСКИХ ТЕКСТОВ, ИНТЕГРИРОВАННАЯ С ИНФОРМАЦИОННО-ПОИСКОВЫМИ СИСТЕМАМИ*

И. В. Ковалев[1], К. В. Полянский[2], П. В. Зеленков[1], В. В. Брезицкая[1], Г. А. Сидорова[1]

[1]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Россия, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31. E-mail: kleniks@yandex.ru
[2]Сибирский федеральный университет
Россия, 660074, Красноярск, ул. Киренского, 26. E-mail: kostyan785@mail.ru

Рассмотрены современные системы машинного перевода и предложена система поиска, анализа и обработки мультилингвистических текстов, интегрированная с информационно-поисковыми системами сети Интернет. Приведена структура системы, определены задачи, которые необходимо решить при ее разработке.

Ключевые слова: машинный перевод, обработка текста, мультилингвистика, информационно-поисковые системы.

SYSTEM OF SEARCH, ANALYSIS AND PROCESSING OF MULTILINGUISTIC TEXTS, INTEGRATED WITH DATA RETRIEVAL SYSTEMS

I. V. Kovalev[1], K. V. Polyanskiy[2], P. V. Zelenkov[1], V. V. Brezitskaya[1], G. A. Sidorova[1]

[1]Siberian State Aerospace University named after academician M. F. Reshetnev
31 “Krasnoyarskiy Rabochiy” prospect, Krasnoyarsk, 660014, Russia. E-mail: kleniks@yandex.ru
[2]Siberian Federal University
26 Kirenskiy street, Krasnoyarsk, 660074, Russia. E-mail: kostyan785@mail.ru

Modern machine translation systems are considered, problems of the given systems are revealed, possible decisions are offered. The comparative analysis of six most popular systems is made, conclusions are drawn, on quality of their translation. The system of search, analysis and processing of multilinguistic texts, integrated with data retrieval systems of network Internet, is considered. The structure of the given system is presented, problems to be solved for its developing, are defined.

Keywords: machine translation, text processing, multilinguistics, information retrieval systems.

Современные системы машинного перевода (СМП) используют один из двух подходов: память переводов и методы машинного перевода. Качество перевода в таких системах отличается, и на сегодняшний день не существует системы машинного перевода, осуществляющей перевод безошибочно. Задача минимизации ошибок систем машинного перевода, а следовательно улучшения качества перевода, является главной. Память переводов и методы машинного перевода несовершенны и приводят к ошибкам. Лучше человека на данный момент переводить не может ни одна система. Поэтому чем больше вовлеченность человека в процесс перевода, тем лучше и качество переведенного текста.

Однако несмотря на это качество перевода современных СМП возрастает как за счет применения моделей и методов искусственного интеллекта, так и за счет использования сервисов сети Интернет. Рассмотрим наиболее актуальные проблемы, с которыми сталкиваются СМП сегодня.

Привязанность к переводческим парам. Большинство современных СМП построено по принципу переводческих пар (рис. 1). Под переводческой парой следует понимать совокупность правил перевода, переводящей грамматики, контекстологических, терминологических, фразеологических и других словарей, а также иных правил и алгоритмов, отвечающих за перевод в направлении данной пары «исходный язык – целевой язык» (ИЯ–ЦЯ).

Подход переводческих пар обусловлен возможностью более узкого описания поведения языка в рамках конкретного направления перевода. Таким образом, каждая переводческая пара – это подробное описание перевода «ИЯ–ЦЯ». Наличие большого количества переводческих пар придает СМП громоздкость, обусловленную большим объемом информации и присутствием избыточности в описаниях. Добавление новой языковой пары требует кропотливой работы и сильно усложняет систему. При возникновении ошибки в грамматике одного из языков необходимо выполнить редактирование в грамматиках языков, составляющих с ним переводческие пары. Наряду с неоспоримыми достоинствами подхода переводческих пар приведенные недостатки являются существенными и создают ряд серьезных проблем.

Решение данных проблем может быть следующим.

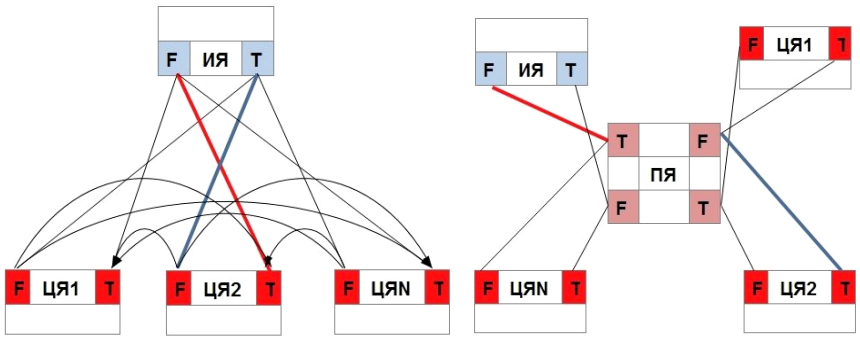
* Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, ГК № 16.740.11.0750.

Внедрение систем Interlingua. Модель СМП с переводческими парами можно представить в виде набора двухзвенных отношений «ИЯ–ЦЯ», где каждое отношение является автономным и не зависит от изменений, вносимых в смежные отношения (рис. 1, 2). Модель СМП, построенных по типу Interlingua, представляет собой архитектуру типа «Звезда» с набором трехзвенных отношений «ИЯ–ПЯ–ЦЯ». Звено ПЯ (промежуточный язык) располагается в центре звезды, в то время как остальные языки – на периферии. В качестве ПЯ может выступать любой язык (обычно английский), для которого строятся переводческие пары «ИЯ–ПЯ» и «ПЯ–ЦЯ», как показано на рис. 1, *б*.

Из рис. 1 видно, что для каждого перевода с ИЯ на ЦЯ в модели «Переводческие пары» задействовано одно отношение «ИЯ–ЦЯ», в то время как модель «Звезда» использует два отношения «ИЯ–ПЯ» и «ПЯ–ЦЯ». Избыточность отношений для одного перевода в модели «Звезда» компенсируется минимальным количеством всех отношений в случае многих языков. Так, если СМП делает перевод на N языков, то количество отношений в схеме «Переводческие пары» равно $N(N-1)$, а количество отношений в схеме «Звезда» определяется как $2N-2$. При большом значении N схема «Звезда» становится менее громоздкой и более простой в процессе разработки СМП, что видно из графика на рис. 2.

За счет наличия ПЯ подход Interlingua также обеспечивает единообразие при переводе на несколько ЦЯ. Это весьма востребованно в последнее время при составлении многоязычной технической документации, а также и руководств пользователя. Подход Interlingua решает проблему переводческих пар.

Использование концептно-ориентированной модели СМП. Еще одной альтернативой использования переводческих пар является концептно-ориентированная модель СМП. Данная модель также построена по типу «Звезда». Однако звено ПЯ, в отличие от подхода Interlingua составляют не ПЯ-термы (термы промежуточного языка), а концепты. Концептом называется не зависящее от конкретного языка понятие, соответствующее реальной или абстрактной сущности, свойству, действию либо иному элементу, отражающему связь между другими понятиями. Концепты берутся из специализированного словаря, на основе которого строится граф концептов. Граф концептов является базой в звене ПЯ концептно-ориентированных СМП. Для построения графа концептов в современных СМП используется язык UNL (The Universal Networking Language) – искусственный семантико-синтаксический язык, не зависящий от какого-либо естественного языка. Наряду с тем что концептно-ориентированные модели решают проблему переводческих пар, они обладают мощным эвристическим инструментом. На основе графа концептов производится довольно качественный семантический анализ, что играет большую роль в осуществлении осмысленного перевода [1].



а *б*

Рис. 1. Модели «Переводческие пары» (*а*) и «Звезда» (*б*)

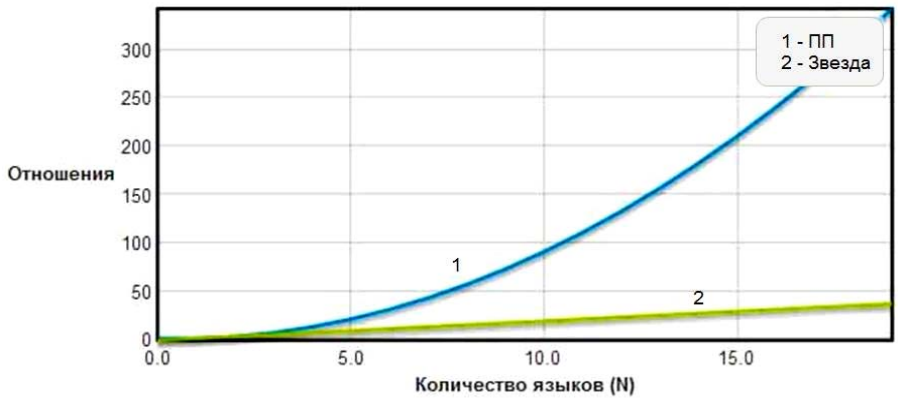


Рис. 2. График зависимости количества отношений от количества языков в моделях «Переводческие пары» (1) и «Звезда» (2)

Отсутствие качественных контекстуальных и семантических анализаторов. Качественный перевод на лексическом и синтаксическом уровне требует учета семантической составляющей текста. Для этого необходим механизм нахождения вариантов перевода на ЦЯ, наилучшим образом подходящих по смыслу, что невозможно сделать без учета контекста. Однако даже для самых современных СМП не удалось разработать более или менее качественных контекстуальных анализаторов. Этим и объясняется наличие в большинстве СМП проблемы, описанной философом Бар-Хиллелом.

Бар-Хиллел выступил с утверждением о принципиальной невозможности осуществления высококачественного полностью автоматического машинного перевода. В качестве примера была приведена проблема отыскания правильного перевода для слова *pen* в следующем тексте: *John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy*. *Pen* в данном случае должно переводиться не как «ручка», «перо» (инструмент для письма) или «загон», а как «детский манеж» (*play-pen*). Выбор того или иного перевода в подобном случае обусловлен знанием внеязыковой действительности, а это знание слишком обширно и разнообразно, чтобы вводить его в компьютер. Таким образом, Бар-Хиллел утверждал, что, не зная контекста, СМП никогда не построит осознанно правильного предложения. Это актуальная и очень острая проблема. Результаты переводов фразы Бар-Хиллела экспертом и современными СМП представлены ниже.

– перевод эксперта: Джон искал свою игрушечную коробку. Наконец он ее нашел. Коробка была в манеже. Джон был очень счастлив;

– Яндекс. Перевод (URL: http://translate.yandex.ru/): Джон смотрел на его коробке. Наконец он нашел его. Коробка была в загоне. Джон был очень счастлив;

– PROMT. Translate (URL: http://translate.ru/): Джон искал свою игрушечную коробку. Наконец он нашел это. Коробка была в ручке. Джон был очень счастлив;

– Google Переводчик (URL: http://translate.google.com/?): Джон искал его игрушки коробке. Наконец, он нашел ее. Коробка была в загоне. Джон был очень счастлив;

– <META> Переводчик (URL: http://translate.meta.ua/): Джон искал его игрушечную коробку. Наконец он нашел это. Коробка была в ручке. Джон был очень счастлив;

– SYSTRANet (URL: http://www.systranet.com/translate): Джон искало его коробка игрушки. Окончательно он нашел он. Коробка находилась в ручке. Джон было очень счастливо;

– Bing Translator (URL: http://microsofttranslator.com/): Джон искал его игрушек окна. Наконец он нашел его. Поле был в перо. Джон был очень счастлив;

– Babylon 9 Онлайн (URL: http://perevodchik.babylon.com/): Джон изучает его игрушки ввода. Наконец, он счел. В графе было в ручку. Джон был очень рад.

Исходя из этих результатов можно сделать следующие выводы:

– качество перевода в приведенных СМП отличается;

– нет СМП, которые бы опровергли утверждение Бар-Хиллела при переводе третьего предложения;

– различные СМП неодинаково ведут себя при переводе одних и тех же предложений, причем ошибки перевода порой не зависят от сложности предложений. Так, например, СМП «Google Переводчик» не справилась с переводом несложного предложения John was looking for his toy box, переведя его как «Джон искал его игрушки коробке», в то время как менее популярная СМП «<META> Переводчик» перевела это предложение почти правильно: «Джон искал его игрушечную коробку». Такое явление можно объяснить разными подходами, используемыми в различных СМП при переводе.

То, насколько качественно рассматриваемые СМП выполнили перевод предложений Бар-Хиллела, показано на рис. 3. В качестве критерия оценки качества перевода было взято процентное соотношение соответствия термов для четырех рассматриваемых ЦЯ-предложений каждой СМП с термами четырех аналогичных ЦЯ-предложений, переведенных экспертом-переводчиком.

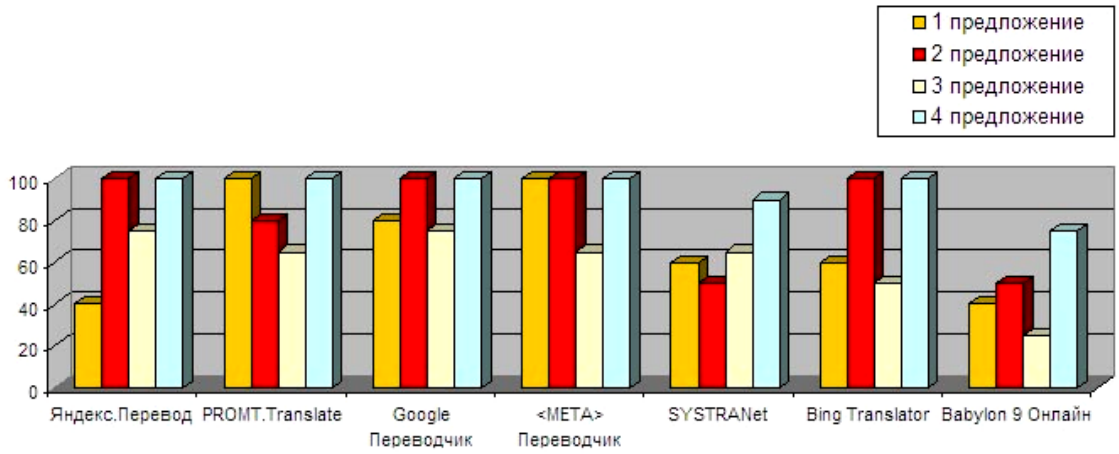


Рис. 3. Гистограмма соответствия переводов экспертному варианту в предложениях Бар-Хиллела

Усредняя результаты, приведенные на рис. 3, по всем СМП, получим следующие показатели качества перевода предложений, приведенные в таблице.

№ п/п	ИЯ-предложение	Качество перевода в среднем, %
1	John was looking for his toy box	68,57
2	Finally he found it	82,86
3	The box was in the pen	60,00
4	John was very happy	95,00

Из таблицы видно, что наиболее качественным для всех СМП является перевод второго и четвертого предложений. Данные предложения – простые по своей структуре, состоят из группы подлежащего и сказуемого и не содержат сложных речевых оборотов. Перевод этих предложений не представляется сложным. Ухудшение качества перевода для первого предложения обусловлено наличием словосочетания *toy box*, перевод которого лексически дифференцирован по типу переводящей СМП. Такое явление связано с отличиями в терминологической базе используемых СМП словарей либо с отличиями в предпочтениях переводящих алгоритмов, которые выбирают конечный вариант перевода из многих возможных. Результаты перевода третьего предложения являются плохими. Несмотря на то что сходство с экспертным переводом для этого предложения составляет в среднем 60 %, у таких СМП, как *Bing Translator* и *Babylon 9* Онлайн, показатель качества не превышает 50 %. В целом ни одна СМП не перевела терм *pen* в соответствии с требованиями контекста. Это, в свою очередь, свидетельствует о неудовлетворительном произведении контекстуального анализа текста и семантическом анализе термов ИЯ-предложения. На сегодня это острая проблема для всех СМП.

Статичность информационных баз. В современных СМП используются такие Информационные базы (ИБ) как словари, базы данных, базы знаний.

Словарь – один из наиболее популярных типов ИБ. Словарями в СМП называются двунаправленные упорядоченные наборы термов или их сочетаний, в которых каждому терму ИЯ ставится в соответствие один или несколько термов ЦЯ. Помимо двуязычных словарей, получивших широкое распространение, существуют и мультиязычные словари [2]. В таких словарях ИЯ-терму могут быть сопоставлены термы нескольких ЦЯ. В случае мультиязычных словарей между термами образуются множественные взаимосвязи, поэтому построение таких словарей – задача сложная, требующая больших временных затрат. Составленный однажды, словарь становится статичной ИБ, претерпевающей незначительные корректировки в длительной временной перспективе. Но поскольку любой язык – сущность, динамически развивающаяся во времени, то изменения, происходящие в языке, должны находить свое отражение и в словаре [3].

Проблема статичности ИБ может быть решена за счет обновления словаря через информационно-поисковые системы (ИПС) сети Интернет. Получение синтаксически корректного, но неточного с лексической точки зрения перевода часто вызвано устареванием термов, составляющих ИБ. В этом случае СМП необходима синхронизация обновлений ИБ с каким-либо источником лингвистической информации. Интернет является подходящим для этой цели средством, ведь терминологическая база глобальной сети постоянно обновляется за счет развития языка составляющих ее статей. Процесс обновления ИБ через сеть Интернет осуществляется посредством ИПС путем выполнения последней информационно-поисковых запросов. Результатом таких запросов являются наборы термов, потенциально пригодных для обновления ИБ. После проверки данных термов на лексическую и морфологическую корректность часть из них может быть предложена к добавлению. Взаимодействие СМП с ИПС сети Интернет решает проблему статичности ИБ. Это приобретает наибольшую актуальность в случае громоздких мультиязычных словарей [4].

Итак, очевидно, что гипертекстовые ресурсы сети Интернет являются самым большим языковым корпусом и содержат большой объем текстовой информации. Более того, гипертекст Интернета полностью создан человеком, за счет чего гипертекст обладает в среднем высоким процентом грамматической правильности. Следовательно, использование данного корпуса в качестве памяти перевода позволит минимизировать ошибки перевода. Для извлечения из сети Интернет лингвистической информации, необходимой при переводе, предлагается использовать информационно-поисковые системы. Полученная система машинного перевода должна принадлежать классу IP-СМП.

Структура IP-СМП. IP-СМП – это система машинного перевода, использующая ИПС сети Интернет для осуществления машинного перевода. IP-СМП обладает рядом свойств, позволяющих эффективно решать поставленные задачи, и является веб-приложением. Распределенность IP-СМП через сеть Интернет позволяет решить такие проблемы, как встраиваемость в службы, сервисы и приложения, доступность онлайн, обеспечение обновления информационной базы.

IP-СМП функционирует следующим образом. Поступивший на вход системы текст на исходном языке должен быть в первую очередь тщательно проанализирован. Целью анализа является формирование на выходе характерных термов, то есть термов, наиболее удачно описывающих предметную область ИЯ-текста. На основе этих термов генерируется и отправляется запрос к ИПС для получения коллекции релевантных документов на целевом языке. Если в классических МТ-системах на стадии поиска цепочек перевода происходит обращение к БД, то IP-СМП обращается за информацией к ИПС сети Интернет. Из полученных от ИПС текстов на ЦЯ будет синтезирован ЦЯ-текст, являющийся переводом данного ИЯ-текста [5].

К IP-СМП предъявляются следующие требования:

– интеграция с ИПС сети Интернет;

– реализация памяти переводов;

– веб-интерфейс.

В соответствии с [6] и указанными требованиями архитектура IP-СМП будет иметь вид, представленный на рис. 4.

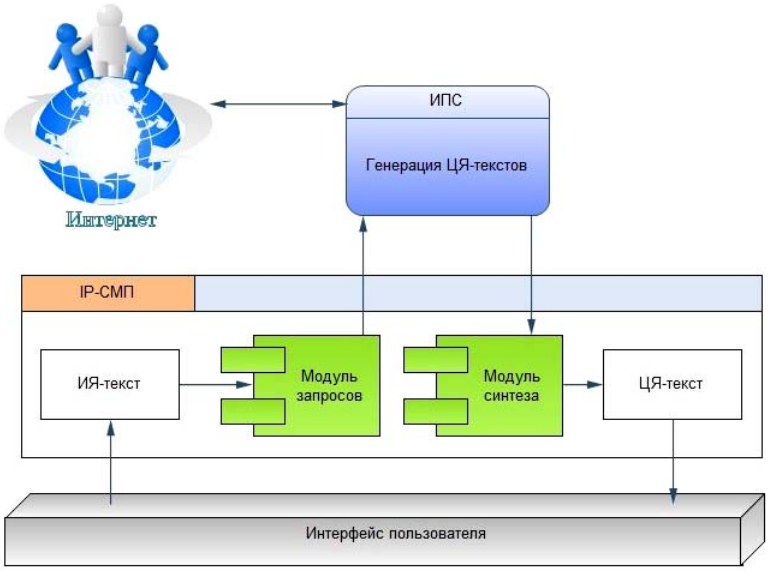


Рис. 4. Архитектура IP-СМП

Основными блоками IP-СМП являются:

– интерфейс пользователя;

– модуль запросов;

– модуль синтеза.

Их взаимодействие осуществляется следующим образом. Полученный через интерфейс пользователя ИЯ-текст подвергается анализу, для чего используются частотные алгоритмы взвешивания термов. Результат анализа поступает на вход модуля запросов в виде характерных термов. Данные термы предварительно переводятся на ЦЯ, и на их основе строится строка поискового запроса, с которой модуль обращается к ИПС. ИПС в ответ на запрос возвращает коллекцию ЦЯ-текстов, являющихся релевантными данному ИЯ-тексту. Для ИЯ и ЦЯ-текстов строится скрытая марковская цепь (СМЦ) перевода. Задача модуля синтеза состоит в генерации перевода на ЦЯ на основе построенной СМЦ. Результат работы модуля поступает в интерфейс пользователя в виде окончательного перевода на ЦЯ [7].

Разработка IP-СМП затрагивает следующие вопросы:

– выбор алгоритмов частотного взвешивания термов;

– взаимодействие системы с ИПС;

– оптимизацию языка поисковых запросов применительно к целям IP-СМП;

– построение СМЦ перевода для ИЯ/ЦЯ-текстов;

– организация памяти переводов;

– оценка качества перевода.

Итак, рассмотрев шесть наиболее популярных систем машинного перевода, можно сделать вывод о том, что все переводы имеют отклонение от экспертного, взятого за эталон. Наличие проблемы Бар-Хиллела наблюдается во всех рассматриваемых СМП. В результате анализа наиболее актуальных проблем современного перевода предложено возможное решение в виде системы поиска, анализа и обработки мультилингвистических текстов, интегрированной с информационно-поисковыми системами сети Интернет, приведена структура данной системы, определены задачи, которые необходимо решить при ее разработке.

Библиографические ссылки

1. *Марчук Ю. Н.* Проблемы машинного перевода. М. : Наука, 1983. С. 12–15.

2. *Ковалев И. В., Усачев А. В.* Мультилингвистический метод изучения иностранной терминологической лексики на базе мнемотехнического подхода // Соц.-психол.пробл. развития личности : материалы I Всерос. науч. интернет-конф. Вып. 4. Тамбов : ТГУ, 2001. С. 58.

3. *Ковалев И. В., Карасева М. В., Лесков В. О.* Информационно-терминологический базис как совокупность лексически связанных компонентов // Вестник СибГАУ. 2009. № 1 (22). С. 54–56.

4. *Kovalev I., Kovaleva T., Susdaleva E.* Effective Information Training Technology Based on the Learner's Memory State Model. Modelling, Measurement and Control D. 2000. Т. 21, № 3–4. С. 11–26.

5. *Карасева М. В., Ковалев И. В., Суздалева Е. А.* Модель архитектуры мультилингвистической адаптивно-обучающей технологии // Новые информационные технологии в университетском образовании : тез. междунар. науч.-метод, конф. Кемерово, 2002. С. 192–195.

6. *Ковалев И. В.* Системная архитектура мультилингвистической адаптивно-обучающей технологии и современная структурная методология // Дистанц. и виртуал. обучение. 2002. № 3. С. 6–12.

7. *Нелюбин Л. Л.* Компьютерная лингвистика и машинный перевод. М.,1991. С. 68–71.