СТАТИСТИЧЕСКАЯ СИСТЕМА МАШИННОГО ПЕРЕВОДА

Л.В. Серебряная, Г.Э. Романюк**, В.С. Демидович**, Д.В. Мардас***

*Белорусский государственный университет информатики и радиоэлектроники,
кафедра программного обеспечения информационных технологий
П.Бровки, 6, 220013, г. Минск, Беларусь
(+375 29)773 95 09, l_silver@mail.ru
**Белорусский Национальный Технический Университет, кафедра интеллектуальных систем
пр-т Независимости, 65, 220000, г. Минск, Беларусь
(+375 17)293 93 25, galarom@tut.by
web: www.bntu.by

Разработана концептуальная модель функционирования системы статистического машинного перевода, принцип работы которой заключается в извлечении информации о переводе из параллельного корпуса заранее переведенных текстов. В соответствии с разработанной моделью создан параллельный корпус текстов и программно реализована система перевода.

Ключевые слова: корпус параллельных текстов, машинный перевод, статистическая модель перевода, триграммная модель.

1 СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД

Стремительные потоки информации, лавина научно-технической документации, получаемая студентами и преподавателями, требуют совершенно нового подхода к проблеме перевода технической литературы. Выход один: максимально автоматизировать процесс, оставив человеку его творческую редакционную часть. В этом помогают системы машинного перевода (МП).

Машинный перевод – процесс перевода текстов (письменных, а в идеале и устных) с одного естественного языка на другой при помощи специализированных компьютерных средств.

Эффективность работы современной системы МП в решающей степени зависит от ее удачной настройки на конкретный подъязык (или микроподъязык) естественного языка, на определенную лексику и ограниченный набор грамматических средств, характерных для текстов данной предметной области, а также на определенные типы документов. Подъязык, с точки зрения МП, определяется в первую очередь некоторым исходным набором текстов. Важную роль играют параллельные тексты и словари-конкордансы, с помощью которых можно достаточно эффективно изучить и использовать в составлении алгоритмов лексическую сочетаемость и дистрибуцию (распределение) языковых элементов в речи или тексте. В любом из современных видов МП необходимо участие человека-редактора, удобство работы которого обеспечивается качеством и надежностью соответствующего программного обеспечения [1].

Наиболее распространенными в настоящее время являются две разновидности машинного перевода: традиционный машинный перевод и статистический машинный перевод.

«Традиционным» называется метод МП, который основан на описании правил и алгоритмов построения человеческой речи.

Достоинством данного метода является ровное качество для любых предложений.

Слабыми сторонами помимо сложности лингвистического разбора предложения и создания правил перевода является также то, что метод требует труда квалифицированных лингвистов для создания моделей каждого языка.

Статистический машинный перевод (СМП) – разновидность МП, которая основана на поиске наиболее вероятного перевода предложения с использованием данных, полученных из двуязычной совокупности текстов (параллельного корпуса) в результате обучения (по языковым парам).

Языковые пары – тексты, содержащие предложения на одном языке и соответствующие им предложения на втором. Языковые пары могут быть как вариантами написания двух предложений человеком – носителем двух языков, так и переводом с исходного на язык перевода, выполненный человеком. Чем большим количеством языковых пар располагает система, и чем точнее они соответствуют друг другу, тем лучшие результат статистического машинного перевода.

В процессе работы система анализирует огромные словарные базы парных фрагментов (фраз из двух-трех слов) – оригинал фрагмента и его перевод. Программа вычисляет наиболее вероятную последовательность слов выходного языка, которую она считает соответствующей переводу исходного текста. В отличие от традиционных систем перевода статистическая программа не учитывает в своей работе грамматические правила. Такую обработку приходится применять к уже переведенному тексту.

Для получения качественного перевода система должна располагать большой базой параллельных переводов.

Сильными сторонами данного метода является то, что нет необходимости вручную создавать правила семантического разбора и перевода предложений. А значит, отпадает необходимость в труде высококвалифицированных лингвистов.

Поэтому для создания СМП был избран именно статистический подход, как более перспективный и не требующий специфических знаний особенностей построения человеческих языков.

При статистическом подходе проблема перевода рассматривается в терминах канала с помехами. Представим себе, что нам нужно перевести предложение с английского на русский. Принцип канала с помехами предлагает нам следующее объяснение отношений между английской и русской фразой: английское предложение представляет собой ни что иное, как русское предложение, искаженное неким шумом.

Для того чтобы восстановить исходное русское предложение, нам нужно знать, что именно люди обычно говорят по-русски и как русские фразы искажаются до состояния английского. Перевод осуществляется путем поиска такого русского предложения, которое максимизирует произведения безусловной вероятности русского предложения и вероятности английского предложения (оригинала) при условии данного русского предложения. Согласно теореме Байеса (формула (2.1)), это русское предложение является наиболее вероятным переводом английского:

$$\arg\max_{e} P(e \mid f) = \arg\max_{e} P(f \mid e) * P(e). \qquad (2.1)$$

где e – предложение перевода;

f – предложение оригинала (источника).

Таким образом, нам требуется модель источника и модель канала, или модель языка и модель перевода. Модель языка ($P(e)$ в формуле 2.1) должна присваивать оценку вероятности любому предложению конечного языка (например, русского), а модель перевода ($P(f|e)$ в формуле 2.1) должна присваивать оценку вероятности предложения оригинала при условии определенного предложения на конечном языке [2].

В общем случае система машинного перевода работает в двух режимах:

1) Обучение: берется тренировочный корпус параллельных текстов, и с помощью линейного программирования ищутся такие значения таблиц переводных соответствий, которые максимизируют вероятность языковой (например, русской) части корпуса при имеющейся английской согласно выбранной модели перевода. На русской части того же корпуса строится модель русского языка.

2) Эксплуатация: на основе полученных данных для незнакомого английского предложения ищется русское, максимизирующее произведение вероятностей, присваиваемых моделью языка и моделью перевода.

Общий алгоритм взаимодействия блоков системы представлен на рисунке 1.

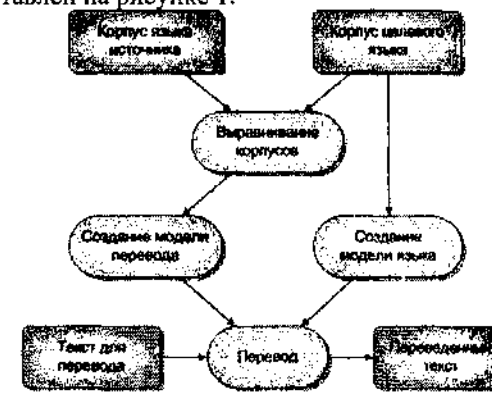

Рис. 1. Общий алгоритм взаимодействия блоков системы

Под корпусами языка источника и целевого языка понимаются огромные базы текстов, насчитывающие сотни тысяч предложений.

Рассмотрим подробнее основные блоки системы.

2 ВЫРАВНИВАНИЕ КОРПУСОВ ТЕКСТОВ

Параллельный текст (битекст) — текст на одном языке вместе с его переводом на другой язык. «Выравнивание параллельного текста» — это идентификация соответствующих друг другу предложений в обеих половинах параллельного текста. Большие собрания параллельных текстов называются «параллельным корпусом» (англ. parallel corpora). Выравнивание параллельного корпуса на уровне предложений является необходимым условием для успешного обучения системы СМП. В процессе перевода предложения могут разделяться, сливаться, удаляться, вставляться или менять последовательность. В связи с этим выравнивание часто становится сложной задачей.

Параллельные тексты создаются с помощью специальных компьютерных программ, которые называются «инструментами для выравнивания» (alignment tool), которые позволяют автоматически выравнивать оригинальную версию текста и его перевод. Подобные программы, как правило, приводят в соответствие два текста (оригинал и перевод) по каждому предложению. Собрание параллельных текстов называется «двуязычным корпусом» [3].

Общая схема выравнивания текстов представлена на рисунке 2.

В данной работе для построения параллельного корпуса используются материалы ассоциации Project Syndicate (www.project-syndicate.org) (международная ассоциация, объединяющая 420 газет в 150 странах мира).

Русский и английский варианты статей были рекурсивно загружены с сайта проекта. В дальнейшем производилась обработка сохраненных html-файлов на предмет удаления ненужных для корпуса html-тэгов и дополнительной информации со страницы при помощи Python-скрипта.

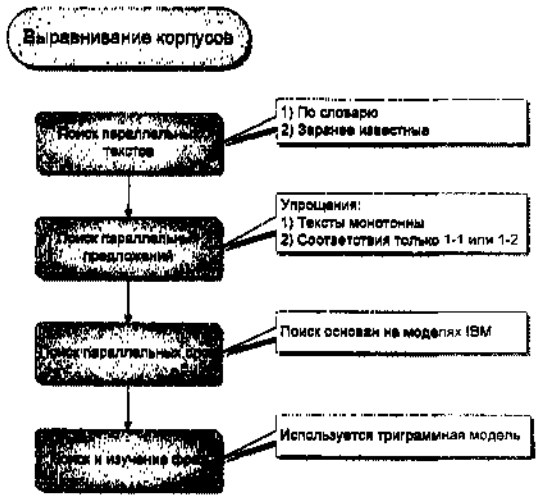

Рис. 2. Схема выравнивания текстов

Для выравнивания полученных текстов необходимо приведение текста к определенному формату. Этот процесс называется токенизацией и сводится к выполнению набора определенных правил форматирования. Такая обработка проводится с помощью отдельного файла с сокращениями и соответствующего perl-скрипта.

Теперь остается выполнить только непосредственное выравнивание предложений. Для этих целей можно воспользоваться различными способами, например, по длине предложения. В этом случае для каждого предложения подсчитывается количество пробелов, и это значение сравнивается с длиной предложения на другом языке.

В результате всех манипуляций получается файл корпуса, в котором в каждой строке следует английское предложение и его русский эквивалент.

Следующим этапом является выравнивание предложений на уровне слов. Для этого осуществляется пословный перевод предложения в обоих направлениях, и находятся общие для обоих вариантов слова. Алгоритм такого процесса представлен на рисунке 3.

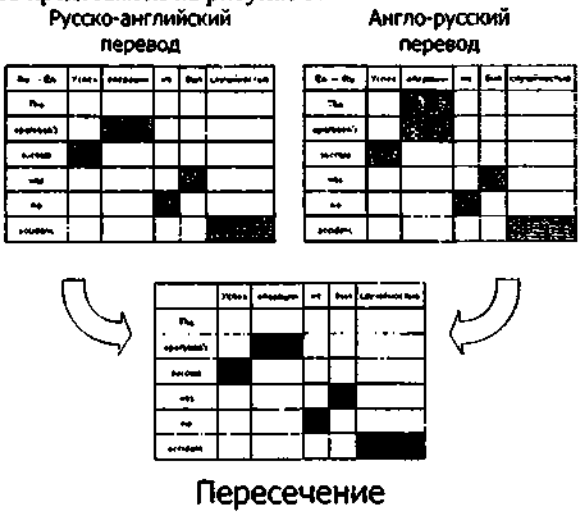

Рис. 3. Выделение слов-пересечений

Для создания таблицы фраз, которая понадобится в последующих модулях, необходимо выделить эти фразы в предложениях. Используется триграммная модель, в которой максимальная длина фразы принимается равной трем словам. Фразой обозначается группа слов, находящихся рядом в таблице переводных соответствий и не содержащая других слов в каждой строке и столбце (рисунок 4).

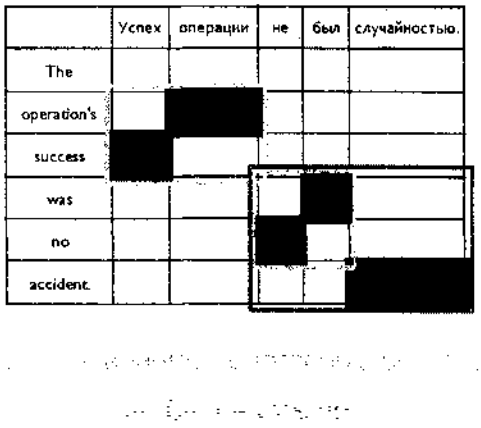

Рис. 4. Выделение фраз в предложении.

Подготовленные таким образом тексты используются для обучения модели перевода.

3 МОДЕЛЬ ПЕРЕВОДА

Самой простой статистической моделью перевода является модель дословного перевода. В этой модели, известной как Модель IBM №1, предполагается, что для перевода предложения с одного языка на другой достаточно перевести все слова (создать «мешок слов»), а расстановку их в правильном порядке обеспечит модель языка. Единственным массивом данных, которым оперирует Модель №1, является таблица вероятностей попарных переводных соответствий слов двух языков.

Обучение Модели №1 производится на корпусе параллельных текстов, выровненном на уровне предложений.

Однако, Модель №1 допускает ситуацию, в которой наиболее употребительным переводом нескольких смысловых слов может быть признано одно высокочастотное – например, служебное – слово конечного языка, и данная модель не всегда правильно учитывает порядок слов в предложении.

Чтобы сохранить при переводе информацию, заключенную в порядке слов, была предложена Модель IBM №2. В этой модели помимо таблицы переводов вводится таблица вероятностей обратных смещений, т.е. вероятностей, что при определенной длине предложения в языке перевода l и длине предложения в языке m оригинала слову перевода в позиции j будет соответствовать слово оригинала в позиции i.

Модель №2 не допускает возможности, что одному слову оригинала соответствует несколько слов перевода. Этот недостаток устраняется в Модели №3, где вводится

понятие коэффициента деления (fertility) слова оригинала и, соответственно, таблица вероятностей каждого значения коэффициента деления для каждого слова.

В Модели №4 и близкой к ней Модели №5 делается следующий шаг к включению понятий грамматики в систему статистического машинного перевода. В Модели №4 появляется понятие класса слов, определяемого автоматически для всех слов языка оригинала и языка перевода.

Обучение моделей №2 – №5 происходит аналогично Модели №1. Так как каждая итерация обучения более сложных моделей занимает существенно больше времени, чем для простых моделей, то обычно перед началом обучения сложных моделей производится несколько итераций младших моделей (начиная с первой), а потом их результаты преобразуются в формат более высоких моделей. Таким образом, оптимизация в старших моделях начинается не со случайного решения, а с некоторого решения, довольно близкого к оптимальному [2].

Для практической реализации метода необходимо произвести следующие подготовительные действия:

- необходимо привести все слова корпусов к нижнему регистру, т.к. для обучения регистр не несет полезной информации
- разбить тексты корпуса на обучающие и тестовые наборы
- создать словарь, который будет содержать сведения о частоте встречаемости слова в корпусе
- заменить своими уникальными идентификаторами все слова в корпусе

Теперь можно приступать к обучению модели. Для обучения используется модуль GIZA [4], входными данными для которого являются словари и файлы выровненных предложений, описанные выше.

Результатом работы модуля являются таблица вероятностей переводов слов и фраз.

4 МОДЕЛЬ ЯЗЫКА

В качестве модели языка в системах статистического перевода используются преимущественно различные модификации n-граммной модели, утверждающей, что грамматичность выбора очередного слова при формировании текста определяется только тем, какие (n–1) слов идут перед ним. Вероятность каждого n-грамма определяется по его встречаемости в тренировочном корпусе.

Самым простым способом моделирования является моделирование в зависимости от контекста (всех предыдущих слов предложения). Однако в целях упрощения этой задачи и из-за требуемых вычислительных ресурсов приходится ограничиваться только некоторым окном. В данной работе используется триграммная модель со сглаживанием, которая оценивает вероятность грамматичности каждого слова Z, следующего в тексте за словами X и Y, по формуле (2).

$$P(Z \mid XY) = \frac{0.95 * F(XYZ)}{F(XY)} + \frac{0.04 * F(YZ)}{F(Z)} + \frac{0.04 * F(Z)}{N} + 0.002 \quad (5.1)$$

где XYZ – фраза из трех слов;
F – частота встречаемости слов
N – общее число слов.

5 ПЕРЕВОД

Перевод осуществляется путем разделения исходного предложения на фразы длиной 3 слова. Под фразами понимается несколько подряд идущих слов без учета их смыслового значения. Для каждой фразы производится поиск наиболее вероятного перевода в таблице фраз таким образом, чтобы перевод всего предложения максимизировал произведения безусловной вероятности переведенного предложения и вероятности предложения оригинала при условии данного переведенного предложения согласно теореме Байеса (формула (2.1)).

Таким образом, модель перевода ($P(f|e)$ будет обеспечивать требуемую точность перевода, а модель языка ($P(e)$ – адекватность и грамматичность переведенного предложения.

Для реализации подхода применяется последний модуль, который занимается непосредственным переводом (он называется декодером). Входными данными для него являются файлы вероятностей перевода и словари, полученные на предыдущих шагах обучения системы.

Интерфейсом между пользователем и системой служит html-страница, содержащая соответствующие поля ввода передающая запросы сервису и выводящая результаты перевода.

ЛИТЕРАТУРА

[1] Марчук, Ю.Н. Проблемы машинного перевода. / Ю.Н. Марчук -- М. 1983. – 232 с.

[2] Рахимбердиев, Б.Н. Эволюция семантики экономической терминологии русского языка в XX веке: дис. кандидата филологических наук / Б.Н. Рахимбердиев. Московский государственный лингвистический университет [Электронный ресурс] – М. 2003. – Режим доступа: http://semantic-evolution.narod.ru/ – Дата доступа: 03.06.2009.

[3] Jahr, M. Whittle: Corpus-Experiment Preparation Tool. / M. Jahr. Summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU)) [Electronic resource] 1999. – Mode of access: http://www.clsp.jhu.edu/ws99/projects/mt/whittle.txt – Date of access: 03.06.2009.

[4] [4] GIZA++: Training of statistical translation models / National Science Foundation Grant No. No. IIS-9820687. [Electronic resource] – 1999. – Mode of access: http://www.fjoch.com/GIZA++.html – Date of access: 03.06.2009.