УДК 004.912

МОДЕЛЬ IP-СИСТЕМЫ МАШИННОГО ПЕРЕВОДА ТЕКСТА НА БАЗЕ ЧАСТОТНОГО ВЗВЕШИВАНИЯ ТЕРМОВ И СКРЫТОЙ МАРКОВСКОЙ ЦЕПИ

К. В. Полянский

Сибирский Федеральный Университет

Поступила в редакцию 22.03.2011 г.

Аннотация. Рассмотрен новый подход машинного перевода – IP-перевод, использующий поисковые системы сети Интернет. Предложены алгоритмы частотного взвешивания термов для генерации релевантных текстов на целевом языке. Модифицирована структура скрытой Марковской цепи применительно к задачам машинного IP-перевода.
Ключевые слова: машинный перевод, частотное взвешивание, скрытая Марковская цепь, обработка текстов.
Annotation. The new approach of machine translation – the IP-translation using search engines of a network the Internet is considered. Algorithms of frequency weighing of terms for generation of relevant texts in target language are offered. The structure of the latent Markov chain with reference to problems of machine IP-transfer is modified.
Key words: machine translation, the frequency weighing, the latent Markov chain, processing of texts.

ВВЕДЕНИЕ

На производительность современных систем машинного перевода (СМП) влияет несколько факторов. К ним можно отнести достижения в системах искусственного интеллекта (СИИ), а также стремительное развитие информационно-поисковых систем (ИПС). Достижения СИИ применяются преимущественно в МТ-системах (системах машинного перевода, использующих наборы переводящих грамматик). Другой класс систем образуют ТМ-системы (системы машинного перевода, использующие память перевода). Под памятью перевода в таких системах обычно понимают хранящиеся в базе данных (БД) цепочки ранее осуществленных переводов, на основе которых строится вывод предложений целевого языка (ЦЯ). Данная статья рассматривает идею об использовании в ТМ-системах сети Интернет в качестве памяти перевода, так как именно Интернет на сегодняшний день является самым обширным электронным источником лингвистической информации. Другое дело, что информация эта еще не структурирована и мало пригодна для целей перевода. И именно поэтому стремительное развитие ИПС, способных к обработке такой информации, может повлиять как на развитие ТМ-систем в целом, так и на развтие IP-систем в частности. IP-системами машинного перевода назовем разновидность ТМ-систем, использующих для перевода ИПС сети Интернет. Рассмотрим механизм IP-перевода.

IP-перевод включает в себя следующие основные процессы:

1. Анализ текста на ИЯ
2. ИП релевантных текстов на ЦЯ
3. Синтез текста на ЦЯ

1.АНАЛИЗ ТЕКСТА НА ИЯ

Поступивший на вход IP-СМП текст на исходном языке (ИЯ) 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 будет сгенерирован и отправлен запрос к ИПС для получения коллекции релевантных документов на ЦЯ. 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.

Анализ ИЯ-текста включает в себя процедуры удаления стоп-термов (знаков пунктуации, служебных слов и т.п.), стемминг (выде-

ление основ термов и приведение схожих основ к одной) и взвешивание. На этапе удаления стоп-термов и стемминга происходит значительное снижение размерности текста.

Процедура взвешивания термов позволяет выделить семантически наиболее значимые для данного текста термы и является частотной, так как присваивание весов для термов осуществляется на основе частоты их встречаемости.

Существует несколько алгоритмов для вычисления локального веса терма. Самыми простыми являются булево взвешивание (b-взвешивание) и взвешивание по частоте терма в тексте (t-взвешивание). Данные типы взвешивания не в полной мере подходит для целей анализа текста. Более пригодным является взвешивание с нормализацией по частоте (c-взвешивание):

$$t_{ij} = 0,5\chi(f_{ij}) + 0,5\left(\frac{f_{ij}}{\max_k f_{kj}}\right), \qquad (1)$$

$$\chi(t) = \begin{cases} 1 \text{ при } t > 0; \\ 0 \text{ при } t = 0. \end{cases},$$

где f_{ij} – частота встречаемости i-го терма в j-ом документе.

В зависимости от частоты встречаемости терма веса при таком взвешивании находятся в интервале [0; 0,5]. Однако разброс в значениях весов термов при таком взвешивании велик. Следующие два типа взвешивания являются логарифмическими и применяются для сглаживания разницы частот встречаемости термов. Логарифмическое взвешивание (l-взвешивание) описывает формула:

$$t_{ij} = \log(f_{ij} + 1). \qquad (2)$$

Альтернативное логарифмическое взвешиване (a-взвешивание) является модификацией предыдущего типа взвешивания и имеет вид:

$$t_{ij} = \chi(f_{ij}) \cdot (\log(f_{ij}) + 1). \qquad (3)$$

Сравнительная характеристика типов взвешивания термов представлена на графике Рисунка 1 для частот термов от 0 до 100. Она показала, что последние два типа взвешивания являются наиболее приемлемыми. В качестве рабочего алгоритма для задачи анализа ИЯ-текста выбрано a-взвешивание [1].

2. ИП РЕЛЕВАНТНЫХ ТЕКСТОВ НА ЦЯ

В ходе проведения анализа ИЯ-текста в результате взвешивания были выявлены характерные термы. Эти термы имеют наибольший вес, а также характеризуют исходный документ с семантической стороны. Группа характерных термов передает смысловую составляющую ИЯ-текста и используется для построения запроса к ИПС с целью получения релевантных текстов на ЦЯ.

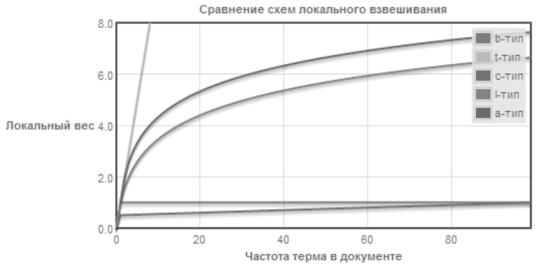

Рис. 1. Сравнение схем локального взвешивания термов

Перед построением запроса к ИПС характерные термы должны быть переведены на ЦЯ. Механизм перевода выполняется посредством имеющихся в системе двунаправленных словарей, имеющих структуру «ИЯ-терм – ЦЯ-терм(ы)». Задача ИП будет заключаться в нахождении таких документов на ЦЯ, для которых подготовленные ЦЯ-термы также будут являться характерными.

В качестве ИПС, осуществляющей поиск релевантных текстов, рассмотрим ИПС «Яндекс», зарекомендовавшую себя при работе с документами на таких языках, как русский, английский, немецкий, французский, итальянский, испанский, украинский и казахский. Язык поисковых запросов ИПС «Яндекс» достаточно велик, и при построении запросов для поиска релевантных ЦЯ-текстов понадобятся лишь некоторые конструкции этого языка, приведенные в Таблице 1.

Так при помощи конструкции lang:langtype можно осуществлять поиск текстов только на ЦЯ, где ЦЯ=langtype, что повышает качество поиска и сужает его область до границ выбранного целевого языка. Конструкции типа «&» и «|» позволяют выполнять строгий и нестрогий типы поиска, когда все либо некоторые термы запроса должны содержаться в предложениях ЦЯ-текста. Таким образом, варьируя приведенные параметры, можно конструировать запросы различных типов. В общем виде запрос к ИПС «Яндекс», основанный на результатах взвешивания, представлен в формуле 4.

Таблица 1

Конструкции языка поисковых запросов ИПС «Яндекс»

Пример	Значение
lang:langtype	Поиск с ограничением по языку langtype
term1 & term2	Термы term1 и term2 в пределах одного предложения
term1 term2 term3 * term5”	Пропущен терм * в выражении
term1 | term2	Поиск любого из термов term1 или term2
term1 ~ term2	Поиск предложения, где терм term1 встречается без терма term2
!!term1	Терм term1 в словарной форме

$$Query = "([term1] \& \&[term2] \& \&... \& \& [termN])lang : [langtype]mime : [mimetype]", \quad (4)$$

где $[term1,...,termN]$ – характерные термы ИЯ-текста, $langtype = ru, en, de, kz$ и т.д. – язык релевантных ЦЯ-текстов, $mimetype = doc, pdf, txt, html$ и т. д. – тип ЦЯ-текстов, получаемых ИПС.

Данный тип запроса наиболее приемлем для обращения к ИПС и позволяет получить на выходе ЦЯ-тексты с хорошей релевантностью. Схема получения релевантных ЦЯ-текстов представлена на рисунке 2.

3.СИНТЕЗ ТЕКСТА НА ЦЯ

После обработки ИЯ-текста и релевантных ему ЦЯ-текстов за счет удаления стоп-термов и стемминга была снижена их размерность. Полученные таким образом тексты помещаются для дальнейшей обработки в линейные списки термов, как показано на Рисунке 3. Затем анализу последовательно подвергается каждое предложение $sentA$ ИЯ-текста. Величиной T обозначим количество термов в $sentA$. Для каждого терма $termA_1,...,termA_T$ строится список его возможных переводов $B_1,...,B_T$ на ЦЯ согласно имеющемуся в системе словарю. Каждый элемент такого списка может содержать различное количество термов-переводов $termB_{i1},...,termB_{iH}$, где величина H зависит от количества переводов терма $termA_i$ на целевой язык. Затем выполняется сканирование текстов $ЦЯ_1,...,ЦЯ_\beta$ на наличие термов из списка возможных переводов, и в случае, если два или более термов в одном ЦЯ-предложении образуют последовательные цепочки, то между термами-переводами (например, $termB_{11}$ и $termB_{2l}$, как показано на рисунке 3) ставится связь. Количество таких последовательно связанных термов-переводов обозначим, как N. Таким образом формируется скрытая Марковская цепь (СМЦ) из N элементов.

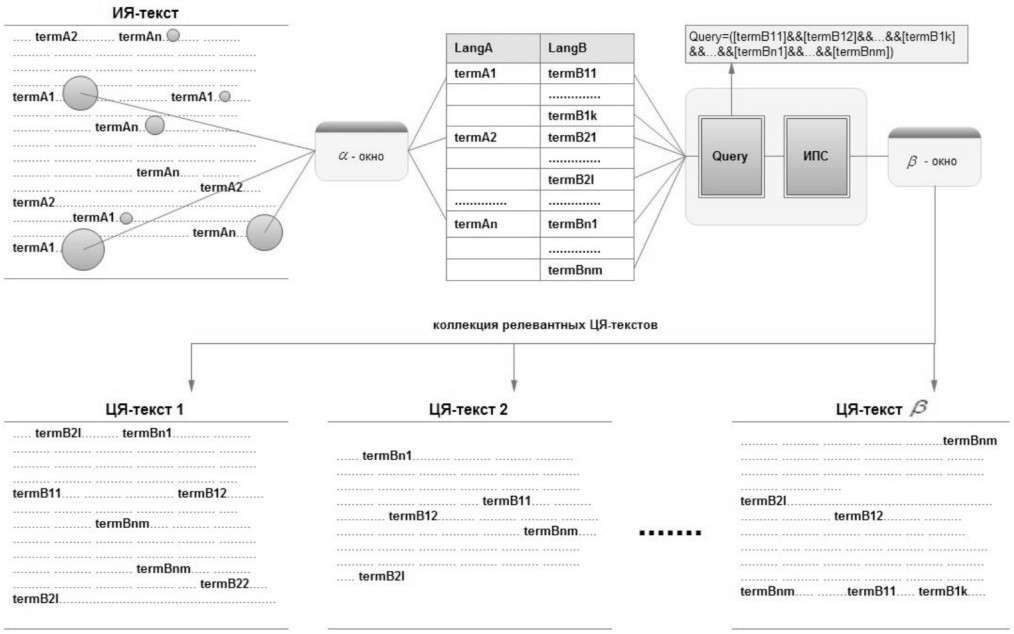

Рис. 2. Механизм получения релевантных ЦЯ-текстов

ОБОЗНАЧЕНИЯ СМЦ

• $A = \{A_{s_i(r)}, ..., A_{s_N(r)}\}$ – цепочка термов ИЯ-текста, где $A_{s_i(r)}$ – i-ый элемент цепочки с порядковым номером в предложении, равным $S_i(r)$, а $r \in \{1, ..., T\}$. Функция $S_i(r)$ принимает значение из этого интервала, соответствующее порядковому номеру терма в рассматриваемом предложении ИЯ-текста. Длина предложения – T термов.

• $B = \{B_{s_1(r)}, ..., B_{s_N(r)}\}$ – цепочка возможных переводов термов ИЯ-текста на ЦЯ согласно имеющемуся в системе словарю.

• $B_{s_i(r)} = \{B_{s_i(r),1}, ..., B_{s_i(r),L(i)}\}$ – термы-переводы для $S_i(r)$ -го элемента из цепочка возможных переводов B, где величина $L(i)$ зависит от данного элемента $B_{s_i(r)}$. В случае однозначного перевода $B_{s_i(r),L(i)} = 0$.

• $p(A_n \mid B_{n,a})$ – апостериорная вероятность того, что терм A_n ИЯ-текста переводится, как терм $B_{n,a}$ ЦЯ-текста в переводе B_n

• $p(B_{n,a} \mid B_{n,b})$ – апостериорная вероятность того, что в ЦЯ-тексте за термом $B_{n,b}$ следует терм $B_{n,a}$

Вероятность $p(A_n \mid B_{n,a})$ задается как отношение количества раз, когда терм A_n был переведен термом $B_{n,a}$ к общему количеству переводов терма A_n на любой из термов, принадлежащих B_n.

Вероятность $p(B_{n,a} \mid B_{n,b})$ задается как отношение количества раз, когда последовательность термов $B_{n,b}B_{n,a}$ была принята в качестве перевода к общему количеству появлений данной последовательности в ЦЯ-текстах [2].

Таким образом, в результате анализа предложения *sentA* ИЯ-текста может быть сформирован набор цепочек вида $\{[A^{[2,4,5]} \mid B^{[1,3,1]}], [A^{[2,4,5]} \mid B^{[1,3,2]}], [A^{[2,4,5]} \mid B^{[2,3,2]}]...\}$, среди которых необходимо выбрать наилучший перевод в релевантных ЦЯ-текстах. Выбор наиболее подходящей цепочки осуществляется по Формуле 5.

$$p(A^{[s_1(r),...,s_N(r)]} \mid B^{[k(s_1(r)),...,k(s_N(r))]}) = \\ = p(B_{s_1(r),k(s_1(r))}) + \\ + \prod_{n=2}^{N} p(B_{s_n(r),k(s_n(r))} \mid B_{s_{n-1}(r),k(s_{n-1}(r))}) \times \\ \times \prod_{n=1}^{N} p(A_{s_n(r)} \mid B_{s_n(r),k(s_n(r))}), \quad (5)$$

где $k(S_i(r))$ – порядковый номер варианта перевода в элементе цепочки возможных переводов $B_{s_i(r)}$, принимающий значения в интервале $\left[1, ..., L(i)\right]$.

В результате, цепочки, получившие наибольшее значение вероятности по формуле 5, будут предложены как наиболее подходящий

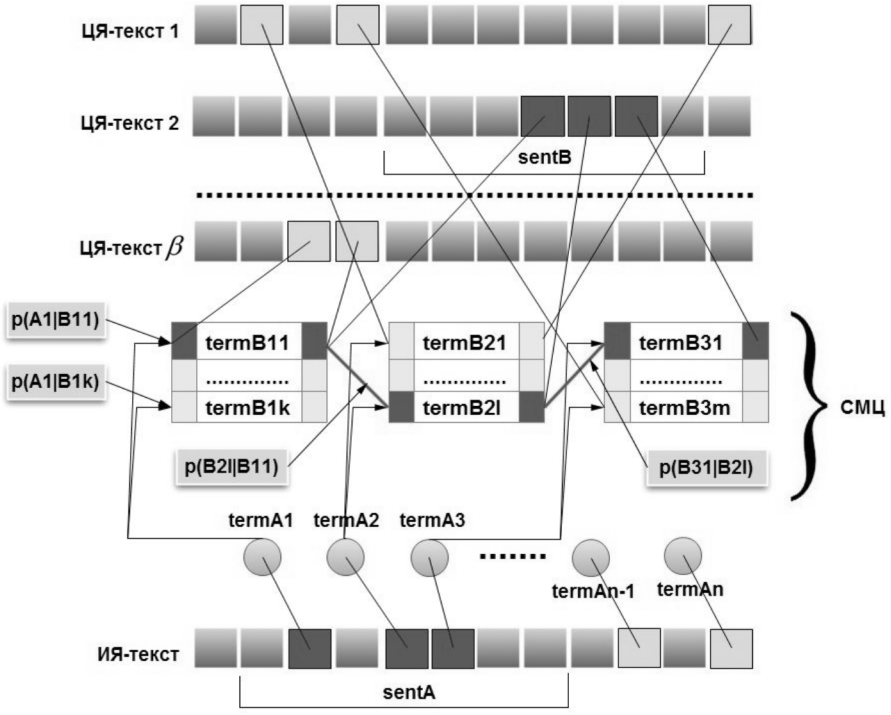

Рис. 3. Скрытая Марковская цепь термов при переводе

вариант перевода на ЦЯ для рассматриваемых термов $A = \{A_{s_1(r)}, ..., A_{s_N(r)}\}$ предложения *sentA* . Далее таким образом переводятся и все остальные предложения ИЯ-текста. Правильность порядка слов в ЦЯ-переводах будет соответствовать правильности порядка слов в сгенерированных на этапе ИП релевантных ЦЯ-текстах. Часть термов, не обнаруженных в ЦЯ-текстах при ИП, переводится на ЦЯ в соответствии с системным словарем и значением вероятности $p(A_n \mid B_{n,a})$ для рассматриваемого терма. После этого можно считать, что перевод на ЦЯ осуществлен.

ВЫВОД

Рассмотренная IP-СМП имеет ряд преимуществ перед СМП, построенными по принципу МТ- и ТМ-систем. Источником информации для выполнения перевода в IP-системе является сеть Интернет, что обеспечивает наличие большого количества потенциальных переводов на ЦЯ. Наличие нескольких алгоритмов взвешивания термов позволяет системе адаптироваться к типу текста и его объему, а также выбирать нужный тип взвешивания при различных условиях. Использование СМЦ для нахождения возможных переводов обеспечивает подбор цепочек терминов с правильным для ЦЯ порядком слов.

СПИСОК ЛИТЕРАТУРЫ

1. *Заболеева-Зотова А. В.* Лингвистическое обеспечение автоматизированных систем / А. В. Заболеева-Зотова, В. А. Камаев // Москва «Высшая школа» – 2008: С. 135–139.

2. *Рубан А. И.* Теория вероятностей и математическая статистика : учеб. пособие / А. И. Рубан. – Красноярск ИПЦ КГТУ. – 2006. – С. 217–219.

Полянский К. В. – аспирант кафедры информатика. Сибирский Федеральный Университет. Тел.:(902) 917-23-00. E-mail: kostyan785@mail.ru

Polyansky K.V. – aspirant of faculty of informatics. Siberian Federal University. Тел.:(902)917-23-00. E-mail: kostyan785@mail.ru