Оценочная функция как средство оптимизации качества МП

С.В. Протасов
Московский Физико-Технический Институт
Россия, 141700, Московская обл., г. Долгопрудный, Институтский пер., д. 9
dialog@svp.zuzino.net.ru

Ключевые слова: машинный перевод, статистический метод, статистический анализ

В распространенных сегодня системах машинного перевода используется словарь переводов слов один к одному, то есть для слова (или словосочетания) на английском (например) языке - сопоставляется одно слово (или словосочетание) на русском. К какой потере смысла это может привести, можно убедиться, переведя текст два раза - туда и обратно. Для улучшение качества перевода обычно предлагается использовать тематические и пользовательские словари, настраиваемые пользователем. Однако число таких словарей ограничено и довольно часто встречается ситуация, что для данной предметной области словари отсутствуют. Для решения этой проблемы в данной статье предлагается:
1. Отказаться от взаимно однозначного сопоставления
2. Поручить системе машинного перевода генерировать много вариантов перевода для каждого предложения
3. Ввести новую оценочную функцию, которая позволяет выбрать оптимальный вариант перевода.
В работе рассматривается детальное описание оценочной функции, конкретные примеры ее применения, сравнение результатов с распространенными системами машинного перевода.

В 90-х годах, когда разрабатывались базовые алгоритмы современных систем машинного перевода компьютерные системы могли себе позволить лишь ограниченные ресурсы и мощности. На жесткий диск максимум размещался один словарь из 10-30 тысяч слов, ни о каком хранении ассоциативного знания между понятиями языка не могло идти и речи. Но на сегодняшний день, когда никого не удивишь сервером с 120 Гигабайтным винчестером и гигабайтом оперативной памяти, полагаю, пришло время для более широких алгоритмов машинного перевода. Могли ли рассчитывать создатели алгоритмов автоматизированного перевода текстов на такие ресурсы лет 8-10 назад? К тому же в наше время Интернет предоставляет уникальные возможности по статистическим исследованиям русского языка, как впрочем и других языков. К примеру Яндекс свободно владеет морфологией русского языка, (т.е. пошел, идти, ходить, ушла - для него одна словоформа), уникальным языком запросов (поиск по расстоянию в словах, в пределах предложения и текста). Как это можно применить для задачи машинного перевода? Можно ли подойти к задаче с другой стороны – оценить, как собственно можно использовать такие ресурсы?

Для начала необходимо выявить типичные ошибки, совершаемые современными системами МП. Условно система МП состоит из нескольких этапов:
1. Морфологический разбор слов предложения на исходном языке.
2. Синтаксический разбор предложения
3. Перевод согласно общему, тематическому словарям (замена слов один к одному)
4. Синтаксическое согласование предложения на конечном языке (перестановка слов).
5. Морфологическое согласование частей речи.

Первое, что бросается в глаза, используя системы, подобные translate.ru, однозначная и часто неверная трактовка перевода слова, совершенная на 3-ем этапе. Попытаемся устранить или уменьшить эту проблему. Будем выбирать те варианты перевода словосочетаний, которые имеют хорошую сочетаемость на большом корпусе текстов. Одной из оценок хорошей сочетаемости может являться количество найденных документов данной комбинации слов. Другими словами, по информации о частотности слов, их взаимной частотности в пределах предложения или частотности по заданному расстоянию в словах можно делать некий вывод о том, насколько часто данные комбинации слов встречаются в языке и выбрать более качественный вариант перевода. Пример:

Таблица 1. Варианты переводов слов предложения и их частотности.

Слово	Перевод	Частотность
experts	эксперты знатоки специалисты	1584160 305199 4490475
are predicting	предсказывают	271618
that		
viruses	вирусы	1628180
and	и	
their	их	
cousins	двоюродные братья кузины сородичи	55165 2743175 48721
the self-propagating	само распространяющиеся	68728
worms	черви	458258
will find	найдут обретут застанут	18403978 471383 181837
new	новые свежие	83965458 1955801
and	и	
even	даже ровно	
more	еще более	
nasty	противный неприятный грязный злобный	526778 579829 579572 286778
ways to	пути дороги способы методы	6643126 9741924 3737897 7671739

	средства	21443990
attack	атаковать нападать на набрасываться на	
computer systems	компьютерные системы	

Введем несколько понятий для заданного корпуса текстов. В качестве корпуса текстов возьмём все документы проиндексированные российской поисковой системой Яндекс. Частотность слова f1 – для данной нормальной формы слова, это число всех словоформ в корпусе текстов. Взаимная частотность fa – это количество документов, в которых присутствуют два заданных слова в произвольной форме на расстоянии в одно слово, то есть рядом, в фиксированном порядке. Нормальная взаимная частотность fx – взаимная частотность, умноженная на число документов в корпусе текстов и деленная на частотность каждого слова. Теперь давайте попробуем выбрать перевод для грамматически связанной пары (viruses, will find). Взаимные частотности (по Яндексу - к-во найденных документов):

Таблица 2. Взаимные частотности слов.

		f1	f2:	fa:	Fx:
червь	найти	10453	348295	31578	573
червь	обрести	10453	348295	40255	288
червь	застать	10453	348295	16468	354

Это может означать, что вариант перевода "черви найдут" , имеющий более высокую взаимную нормальную частотность, более благоприятный, чем "черви застанут" с точки зрения данного метода.

Продолжим, и составим таблицу для всех возможных пар в грамматических конструкциях и для всех вариантов перевода:

Таблица 3. Взаимные частотности слов.

		f1	f2:	fa:	fx:
найти	путь	31578	19250212	56837	723
найти	дорога	31578	19250212	77511	624
найти	способ	31578	19250212	92993	707
найти	метод	31578	19250212	25102	645
найти	средство	31578	19250212	93328	640
обрести	путь	40255	509100	56837	619
обрести	дорога	40255	509100	77511	572
обрести	способ	40255	509100	92993	549
обрести	метод	40255	509100	25102	434
обрести	средство	40255	509100	93328	566
застать	путь	16468	196541	56837	567
застать	дорога	16468	196541	77511	543
застать	способ	16468	196541	92993	496
застать	метод	16468	196541	25102	385
застать	средство	16468	196541	25102	454

Зададим оценочную функцию **f** как сумму всех нормальных взаимных частотностей для всех грамматических пар в данном варианте перевода.

В следующей таблице показано значение оценочной функции **f** и список слов, задающий вариант перевода, для самых высоких значений оценочной функции.

Таблица 4. Значение оценочной функции.

F									
946.50	вирус	червь	найти	нападать	неприятный	путь	эксперт	предска-зывать	сородич
946.54	вирус	червь	найти	нападать	грязный	путь	знаток	предска-зывать	родствен-ник
946.63	вирус	червь	найти	нападать	грязный	путь	эксперт	предска-зывать	сородич
946.79	вирус	червь	найти	нападать	противный	путь	знаток	предска-зывать	родствен-ник
946.88	вирус	червь	найти	нападать	противный	путь	эксперт	предска-зывать	сородич
947.71	вирус	червь	найти	атаковать	неприятный	способ	знаток	предска-зывать	сородич
952.85	вирус	червь	найти	нападать	неприятный	путь	знаток	предска-зывать	сородич
952.97	вирус	червь	найти	нападать	грязный	путь	знаток	предска-зывать	сородич

Таким образом, при данной оценочной функции, оптимальным будет следующий вариант перевода:

Знатоки предсказывают, что вирусы и их сородичи, саморазмножающиеся черви, найдут новые и даже более грязные пути напасть на компьютерные системы.

Сравните с переводом **translate.ru** (Интернет словарь):

Эксперты предсказывают, что вирусы и их кузены, само пропагандирующиеся черви, найдут новые и даже более противные способы напасть на компьютерные системы.

На субъективный взгляд автора первый вариант значительно лучше, несмотря на то, что там вместо "эксперт" используется вариант перевода - "знаток", однако следует отметить, что алгоритм не имеет абсолютно никаких данных о тематике перевода, а данное предложение было выбрано случайным образом и не подгонялось под алгоритм. Хорошо видно, что выбрано немного более удачные и широкие словосочетания “найти путь”, а не “найти способ”, “грязный путь” выиграл у “противный путь”, а «вирус сородич” выиграл у “вирус родственник”. Таким образом, не владея семантическими словарями, данный алгоритм позволил при помощи взаимных частотностей слов позволил “выудить” некоторые удачные семантические сочетания.

Вид оценочной функции (сумма независимых друг от друга оценок) позволяет значительно сократить дерево перебора, отсекая значительную часть вариантов методом ветвей и границ. Это немаловажный фактор для систем машинного перевода, учитывая, что число вариантов перевода для предложения из 15-20 слов может достигать нескольких тысяч.

Для улучшения оценочной функции можно рассматривать не взаимные частотности всех произвольных словоформ, а взаимные частотности конкретных форм слова. Рассмотрим классический пример.

John was looking for his toy box. Finally he found it. **The box was in the pen**. John was very happy. (Джон искал свою игрушечную коробку. Наконец он ее нашел. **Коробка была в манеже**. Джон был очень счастлив.).

Pen в данном случае должно переводиться не как "ручка" (инструмент для письма), а как "детский манеж". Взаимная частотность словосочетания “была в ручке” гораздо меньше, чем “была в манеже”, что позволяет выбрать правильный вариант перевода. Причем в данном примере более важно то, что “ручка” в предложном падеже встречается намного реже, чем “манеж” в предложном и это свойство, которое также может быть автоматически вытянуто из большого корпуса текстов, влияет решительным образом на выбор варианта перевода. (Вариант translate.ru, общий словарь: Джон искал его игрушечную коробку. Наконец он нашел это. **Коробка была в ручке**. Джон был очень счастлив.)

Оценим количество ресурсов необходимых для работы алгоритма. Если взять словарь из 50 тыс. слов, то для хранения всех их взаимных частотностей без оптимизации потребуется всего 2.5 ггбайт памяти на жестком диске, что по состоянию на 2003 год не является чем-то сверхтребовательным. Взаимные частотности всех индивидуальных словоформ, требуют на порядки больше ресурсов, однако и здесь возможны оптимизационные алгоритмы. В частности можно предложить запоминание самых частотных словоформ для каждой нормальной формы, а также самых взаимных частотных словоформ для каждой пары нормальных форм слов.

Предложенный алгоритм сочетает прозрачность, скорость и простоту реализации, не требует создания тематических словарей и не требует человеческих трудозатрат. Эти свойства дают ему хорошие шансы получить широкое распространение в системах машинного перевода следующего поколения.

1. В. А. Дёмин. Корпусная оценка сочетаемости слов с использованием бинарных аналогий. // Труды Международного семинара Диалог ‘2002 по компьютерной лингвистике и ее приложениям. – 2002 г.
2. И. А. Большаков. Какие словосочетания следует хранить в словарях? // Труды Международного семинара Диалог ‘2002 по компьютерной лингвистике и ее приложениям. – 2002 г.
3. Ciaramita M., Johnson M. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. Proceedings of the 18th International Conference on Computational Linguistics, 2000 . Vol. 1. http://arxiv.org/ps/cs/0008020
4. Pereira F., Tishby N., Lee L. Distributional Clustering of English Words // In Proceedings of the "Meeting of the Association for Computational Linguistics", 1993, pp.183-190. http://citeseer.nj.nec.com/article/pereira93distributional.html

5. А. Г. Глазунов Концептно-ориентированная модель памяти переводов http://www.citforum.ru/programming/digest/cotm.shtml
6. Перевод и лингвистика текста/ ред. совет И.И. Убин и др. - М., 1994
7. Марчук Ю.Н. Проблемы машинного перевода. - М.: Наука,1983
8. Mel'cuk, I., A. Zholkovsky. The explanatory combinatorial dictionary // M. Evens (ed.) Relational models of lexicon. Cambridge, England: Cambridge University Press, 1988, P. 41-74.

Key words: machine translation, statistical method, statistical analysis

Evaluation function as means of translation quality optimization.

/ S. V. Protasov (Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, Moscow region, 141700, Russia, dialog@svp.zuzino.net.ru)

Now the usual approach of most modern machine translation (MT) systems is based on using their translation dictionary in one-to-one mode. It means that one English word (or phrase) corresponds to one Russian word or collocation. The usual example of to-and-back translation shows the significant loss of sense as the basic disadvantage of this approach.
The usual solution of this problem is using thematic or other user-defined dictionaries to improve the quality of translation. However the number of dictionaries is limited and often it happens that there are no dictionaries for given knowledge domains.
The solution of this problem given in this article is to refuse the one-to-one relation approach.
The given proposal is to charge the machine translation system with generating many versions of each sentence translations and then to introduce new evaluation function which allows to choose the optimal translation among different translation versions. It is considered detailed description of evaluation function, specific examples of its application, comparison of proposed translation results with results of well-known machine translation systems.