ПОБУКВЕННЫЙ СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД МЕЖДУ РОДСТВЕННЫМИ ЯЗЫКАМИ

Анс БЕРЗИНЬ, Рига, Латвия
Латвийский университет

Abstract. *Usually in the field of the statistical machine translation word is taken as the smallest unit for statistical analysis of the parallel corpus of training of a translation program (and subsequently – also translation). For languages with a great number of morphological forms it makes a problem, because a word cannot be translated, if in the parallel training corpus there is no exactly the same form of the given word. Therefore we propose to take advantage of smaller units of the text – letters and their combinations (or phonems – simple and complex) for translating between cognate languages (i.e., languages having certain morphological and lexical similarity).*

Введение

Обычно при статистическом машинном переводе за единицу статистического анализа параллельного корпуса обучения программы-переводчика (и впоследствии – перевода) берётся слово. Для языков с большим количеством морфологических форм это является проблемой, так как слово не переводится, если в параллельном корпусе обучения не встречалась именно такая форма данного слова. Поэтому у нас возникла мысль попытаться воспользоваться более мелкими единицами текста – буквами и их сочетаниями (или фонемами – простыми и сложными) при переводе между родственными языками (т.е., языками, имеющими определённое морфологическое и лексическое сходство).

История вопроса

Хотя мысль о возможности применения статистических методов в системах машинного перевода прозвучала уже в конце 40-х годов XX столетия, соответствующего уровня развития (и по быстродействию, и по обьёму хранимых данных) ЭВМ достигли лишь к концу 80-х. Первой серьёзной заявкой на применение статистического машинного перевода (СМП) была публикация рабочей группы исследовательского центра ИБМ [1], и уже вскоре они представляли результаты применения своей системы *Кандид* [2]. Суть идеи заключается в созда-

нии статистической модели перевода на основе параллельного корпуса обоих языков (содержащей вероятности соответствия слов исходного языка словам языка перевода), а также статистической модели языка на основе корпуса языка перевода (содержащей вероятности следования слов определённому количеству предшествующих слов в данном языке). Наиболее вероятным переводом, очевидно, будет являться такое слово, при котором максимально умножение этих двух вероятностей. Так как в данной схеме за единицу перевода принимались слова, то ещё учитывались вероятности соответствия данного слова нескольким словам, соответствия пустому слову (т.е., удаления), а также переноса слова на другое место в предложении (при СМП по фразам эти данные содержаться в самой модели перевода).

В конце 90-х начали разрабатываться методы статистического машинного перевода по фразам (СМПФ). Первой с таким предложением выступила группа немецких учёных, модель выравнивания которых по сути является СМПФ [3], потом появились и другие разработки. В 2003 году Коэн, Ох и Марку провели обобщение и сравнение этих методов и в какой-то степени расставили точки над «i» в данной области [4]. Суть различия между СМП по словам и СМПФ в том, что модель перевода содержит соответствия словосочетаний (фраз) (по реализации - n-грамм слов) вместо соответствий слов.

Программное обеспечение

В начале 2006 года мы познакомились с датским аспирантом Яко-вом Эльмингом, разрабатывающим англо-датскую систему СМПФ [6]. Для своего эксперимента мы решили воспользоваться рекоммендованным им программным обеспечением, которое создано Филиппом Коэном:

1) программа ngram-count (из пакета ПО SRI Language Modeling Toolkit) создаёт файл n-граммной модели языка;

2) скрипт train-phrase-model.perl (Филипп Коэн), пользуясь программой *GIZA++* (Франц Иосиф Ох) и вспомогательными программами snt2cooc, mkclass, phrase-extract, phrase-score, создаёт модель перевода, т.е., таблицу переводов фраз;

3) раскодировщиком *Фараон* (*Pharaoh*) (Филипп Коэн) [5] проводится перевод.

Эксперимент

Как мы уже писали во введении, при принятии за наиболее мелкую единицу перевода слово, учитывается лексическая и синтаксическая информация, содержащаяся в параллельном корпусе, а морфологическая и фонетическая - теряется. Посему мы решили про-

вести эксперимент, взяв за наиболее мелкую единицу фонему, т.е., её орфографическое обозначение. (Мы, конечно, могли за единицу взять букву, и это бы сути не изменило и на результаты бы существенно не повлияло, но так как фонемы в большей степени содержат информацию о фонетических соответствиях, то мы свой выбор остановили на них.)

Следовало выбрать пару близких языков, и мы остановили свой выбор на латгальском и латышском. Конечно, параллельных корпусов данных языков не существует, поэтому нам пришлось создать небольшой пробный «корпус» самим. В качестве исходника мы взяли записи латгальских народных песен и снабдили их латышскими переводами. В общей сложности мы ввели 1203 слова (т.е., с точки зрения системы СМПФ – предложения), состоящих из более чем 6000 фонем (т.е., «слов»).

p' ī c' i	p ie c i	a r	a r
b ā r' i	b ē* r i	k o nn u	k a nn u
k u m' e ļ e ņ i	k u m e l i ņ i	m a d s	m e* d u s
v y s' i	v i s i	a r	a r
a r	a r	b o l t u	b a l t u
a u z o m	a u z ā m	b' i č' e r' ei t' i	b i ķ e r ī t i
b a r uo j a m' i	b a r uo j a m i	a s	e s
dz' ī d uo t	dz ie d ā t	a t d ū š u	a t d uo š u
m uo k u	m ā k u	t o v u	t a v u
d o n c uo t	d a n c uo t	d ā l u	d ē* l u
m uo k u	m ā k u	k u r u	k u r u
s' ī n a	s ie n u	d' ī n u	d ie n u
p ļ a u t' i	p ļ a u t i	g r y b ā d a m a	g r i b ē* d a m a
n a m uo c' ie j u	n e m ā c ē j u	i z	u z
t' e+	t e	v a c y s	v e* c a s
o r š o n a	a r š a n a	o k y s	a k a s
e c' ie š o n a	e c ē š a n a	s y r ū b a	g r uo d a
g r yu t s	g r ū t i	s' ē+ d	s ē ž
b' e j a	b i j a	r a s n y m	r e* s n ie m
m a ņ	m a n	b yu d' i m	v ai g ie m
dz' ei v uo t	dz ī v uo t	v a r d' i v' e+	v a r d e
o l s	a l u s		

Таб. 1. Фрагмент параллельного «корпуса».

Дабы иметь возможность пользоваться вышеупомянутым программным обеспечением, мы привели наши данные в вид, им воспринимаемый: фонемы отделили пробелами (т.е., как слова), а слова – переводами строк (т.е., как предложения). Далее мы программой ngram-

count на материале латышской половины нашего «корпуса» создали модель языка, а программой train-phrase-model.perl на всём нашем «корпусе» – модель перевода (в отличие от пословного СМПФ, в котором ограничиваются n-граммами в диапазоне от монограмм до триграмм, максимум – до пентаграмм, мы увеличили содержимое наших моделей до дециграмм, чтобы успешно отражать не только фонетические и морфологические, но и лексические соответствия).

После создания моделей мы приступили к переводу. В качестве переводимого материала мы выбирали народные песни, не включённые в наш «корпус», а также отдельные слова, интересующие нас своими особенными морфологическими свойствами или лексической неродственностью.

Оригинал	Перевод системы СМПФ	Правильный перевод
g uo j u	g ā j u	g ā j u
p a	p a	p a
m’ e ž u	m e ž u	m e ž u
r u b u ļ u	r u b a l u	r u b e ņ u
š au d’ ei d a m s	š au d ī d a m s	š au d ī d a m s
k u r	k u r	k u r
t u	t u	t u
ī s’ i	ie s i	ie s i
c’ ī m a	c ie m a	c ie m a
p ui s’ i	p ui s i	p ui s i
m’ i g ļ e ņ ā	m i g l i ņ ā	m i g l i ņ ā
v o k o r ā	v a k a r ā	v a k a r ā
o k a	a k a	a k a
s y r ū b s	g r uo b s	g r uo d s
s y r ū b a	g r uo d a	g r uo d a
dz’ ī d uo t	dz ie d ā t	dz ie d ā t
d o n c uo t	d a n c uo t	d a n c uo t
c yu k a	c ū k a	c ū k a
s v’ ī s š u	s v ie d ī š u	s v ie d ī š u
m’ e s š u	m e d ī š u	m e t ī š u
b r au č u	b r au c u	b r au c u
p r’ ī š k a	p r ie k š a	p r ie k š a
ī t u m	ie t u	ie t u
g a ņ ei t u	g a n ī t u	g a n ī t
dz a r t u	dz e r t	dz e r t
v u š k y s	ai t a s	ai t a s
r u p u c’ s’	k r u p s	k r u p i s

Таб. 2. Фрагмент результатов перевода.

Результаты оказались весьма хорошими. Слова, переводящиеся соответствиями фонетического уровня, были правильны почти во всех случаях, а морфологического и лексического – если встречались в нашем корпусе достаточно часто («достаточно часто» не значит «часто», при благоприятных обстоятельствах это могут быть и 1-2 вхождения) и им не мешал «шум» других, схожих соответствий. Это вполне закономерно – для фонетических соответствий нужен гораздо меньший обьём исходного статистического материала, так как они повторяются гораздо чаще: в принципе, можно предположить, что фонетические соответствия мы своим «корпусом» охватили все, морфологические – большую часть, а лексические – только некоторые.

Во многих случаях соответствия разных уровней совпадают (например, большинство морфологических соответствий выражаются и через фонетические), но в тех случаях, когда они не совпадают, они оказываются взаимно конкурирующими, скажем, часть слова может переводиться либо пофонемно через фонетические соответствия, либо сразу через морфологическое соответствие, или всё слово – либо на лексическом уровне заменяться на совершенно неродственное слово, либо переводиться по частям фонетическими и морфологическими соответствиями. Конечно, правильное решение зависит от конкретного случая, но всё же можно найти среднее, оптимальное соотношение уровней соответствий и подогнать раскодировщик под него путём изменения его весовых параметров (например, изменяя вес модели языка по отношению к весу модели перевода и т.п.).

Наибольшие проблемы создают омонимы и омографы исходного языка, так как при данной схеме контекст не учитывается вовсе. Это возможно решить путём ввода символа, расделяющего слова́, и обьединения слов в предложения; такое двухуровневое деление на слова и на фонемы позволяло бы учитывать и контекст.

Выводы

Главный вывод оптимистичен – наша система перевода оправдала наши ожидания и работает весьма неплохо. Кроме того, большая часть ошибочных переводов таковыми получились из-за неполности (т.е., малого обьёма) нашего экспериментального «корпуса», поэтому есть основания полагать, что, будучи обученной на достаточно большом корпусе, она бы работала с действительно хорошими показателями правильности перевода.

Конечно, возникает вопрос о реальном, «производственном» применении пофонемного (или побуквенного) СМПФ. Ясно, что между

родственными языками он мог бы применяться. Но скорее всего, так как пословный СМПФ не зависит от родственности языков и уже широко применяется, реальная роль побуквенного СМПФ может быть только вспомогательной пословному СМПФ. Мы считаем весьма перспективной разработку двухуровневой системы СМПФ, которая сперва проводит пословный статистический перевод, а потом – побуквенный для слов, которые не удалось перевести пословным способом. Это бы позволило переводить и те слова, частной формы которых в корпусе нет.

Возможно, некоторые читатели задались вопросом: ради чего всё это, если фонетические и морфологические соответствия можно описать вручную соответствующими правилами. Конечно, в каком-то смысле это так. Однако, для каждой пары языков эти правила пришлось бы описывать заново, а статистическая система по сути является языконезависимой. Кроме того, система перевода с описанием правил уже являлась бы не статистической, а смешанной, а сохранение чистоты жанра также немаловажно.

ЛИТЕРАТУРА

1. *Brown, P. F., Cocke, J. , Della Pietra, S. A., Della Pietra, V. J., Jelinek, F. , Lafferty, J. D., Mercer, R.L., Roossin, P. S.* A Statistical Approach to Machine Translation. // Computational Linguistics. – 1990. – Vol. 16. – No. 2. – P. 79-85.

2. *Berger, A. L. , Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., Ures, L.* The Candide System for Machine Translation. // ARPA Human Language Technology Workshop. San Mateo: Morgan Kaufmann Publishers. – 1994. – P. 152-157.

3. *Och, F. J., Tillmann, C., Ney, H.* Improved Alignment Models for Statistical Machine Translation. // Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora. Maryland: University of Maryland. – 1999. – P. 20-28.

4. *Koehn, P., Och, F. J., Marcu, D.* Statistical phrase-based translation. // Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference. – Edmonton, 2003. – P. 127-133.

5. *Koehn, P.* Pharaoh, a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models: User Manual and Description. Los Angeles: University of Southern California, 2003.

6. Jakob Elmings hjemmeside <*http://www.id.cbs.dk/~je/*>