Mining Parenthetical Translations from the Web by Word Alignment

Dekang Lin
Google, Inc.
Mountain View
CA, 94043
lindek@google.com

Shaojun Zhao[†]
University of Rochester
Rochester
NY, 14627
zhao@cs.rochester.edu

Benjamin Van Durme[†]
University of Rochester
Rochester
NY, 14627
vandurme@cs.rochester.edu

Marius Paşca
Google, Inc.
Mountain View
CA, 94043
mars@google.com

Abstract

Documents in languages such as Chinese, Japanese and Korean sometimes annotate terms with their translations in English inside a pair of parentheses. We present a method to extract such translations from a large collection of web documents by building a partially parallel corpus and use a word alignment algorithm to identify the terms being translated. The method is able to generalize across the translations for different terms and can reliably extract translations that occurred only once in the entire web. Our experiment on Chinese web pages produced more than 26 million pairs of translations, which is over two orders of magnitude more than previous results. We show that the addition of the extracted translation pairs as training data provides significant increase in the BLEU score for a statistical machine translation system.

1 Introduction

In natural language documents, a term (word or phrase) is sometimes followed by its translation in another language in a pair of parentheses. We call these **parenthetical translations**. The following examples are from Chinese web pages (we added underlines to indicate what is being translated):

(1) 美国智库布鲁金斯学会（Brookings Institution）专研跨大西洋恐怖主义的美欧中心研究部主任杰若米·夏皮罗（Jeremy Shapiro）却认为，...

(2) 消化性溃疡的症状往往与消化不良（indigestion），胃炎（gastritis）等其他胃部疾病症状相似.

(3) 殊不知美国是不会接受（not going to fly）这一想法的

(4) …当是一次式时，叫线性规划(linear programming).

[†]Contributions made during an internship at Google

The parenthetically translated terms are typically new words, technical terminologies, idioms, products, titles of movies, books, songs, and names of persons, organizations locations, *etc*. Commonly, an author might use such a parenthetical when a given term has no standard translation (or transliteration), and does not appear in conventional dictionaries. That is, an author might expect a term to be an *out-of-vocabulary* item for the target reader, and thus helpfully provides a reference translation *in situ*.

For example, in (1), the name *Shapiro* was transliterated as 夏皮罗. The name has many other transliterations in web documents, such as 夏皮洛, 夏比洛, 夏布洛, 夏皮羅, 沙皮罗, 夏皮若, 夏庇罗, 夏皮諾, 夏畢洛, 夏比羅, 夏比罗, 夏普羅, 夏批羅, 夏批罗, 夏彼羅, 夏彼罗, 夏培洛, 夏卜尔, 夏匹若 ..., where the three Chinese characters corresponds to the three syllables in Sha-pi-ro respectively. Each syllable may be mapped into different characters: 'Sha' into 夏 or 沙, 'pi' into 皮, 比, 批, and 'ro' into 罗, 洛, 若, ...

Variation is not limited to the effects of phonetic similarity. Story titles, for instance, are commonly translated semantically, often leading to a number of translations that have similar meaning, yet differ greatly in lexicographic form. For example, while the movie title *Syriana* is sometimes phonetically transliterated as 辛瑞那, 辛瑞纳, it may also be translated semantically according to the plot of the movie, e.g., 迷中迷 (mystery in mystery), 实录 (real log), 谍对谍 (spy against spy), 油激暗战 (oil-triggered secret war), 叙利亚 (Syria), 迷经 (mystery journey), ...

The parenthetical translations are extremely valuable both as a stand-alone on-line dictionary and as training data for statistical machine translation systems. They provide fresh data (new words) and cover a much wider range of topics than typical parallel training data for statistical machine translation systems.

Proceedings of ACL-08: HLT, pages 994–1002,
Columbus, Ohio, USA, June 2008.

The main contribution of this paper is a method for mining parenthetical translations by treating text snippets containing candidate pairs as a partially parallel corpus and using a word alignment algorithm to establish the correspondences between in-parenthesis and pre-parenthesis words.

This technique allows us to identify translation pairs even if they only appeared once on the entire web. As a result, we were able to obtain 26.7 million Chinese-English translation pairs from web documents in Chinese. This is over two orders of magnitude more than the number of extracted translation pairs in the previously reported results (Cao, *et al.* 2007).

The next section presents an overview of our algorithm, which is then detailed in Sections 3 and 4. We evaluate our results in Section 5 by comparison with bilingually linked Wikipedia titles and by using the extracted pairs as additional training data in a statistical machine translation system.

2 Mining Parenthetical Translations

A parenthetical translation matches the pattern:

(4) $f_1 f_2 ... f_m (e_1 e_2 ... e_n)$

which is a sequence of m non-English words followed by a sequence of n English words in parentheses. In the remainder of the paper, we assume the non-English text is Chinese, but our technique works for other languages as well.

There have been two approaches to finding such parenthetical translations. One is to assume that the English term $e_1 e_2 ... e_n$ is given and use a search engine to retrieve text snippets containing $e_1 e_2 ... e_n$ from predominately non-English web pages (Nagata *et al*, 2001, Kwok *et al*, 2005). Another method (Cao *et al*, 2007) is to go through a non-English corpus and collect all instances that match the parenthetical pattern in (4). We followed the second approach since it does not require a predefined list of English terms and is amendable for extraction at large scale.

In both cases, one can obtain a list of candidate pairs, where the translation of the in-parenthesis terms is a suffix of the pre-parenthesis text. The lengths and frequency counts of the suffixes have been used to determine what is the translation of the in-parenthesis term (Kwok *et al*, 2005). For example, Table 1 lists a set of Chinese segments (with word-to-word translation underneath) that precede the English term *Lower Egypt*. Owing to the frequency with which 下埃及 appears as a candidate, and in varying contexts, one has a good reason to believe下埃及is the correct translation of *Lower Egypt*.

… 下游 地区 为 下 埃及 downstream region is down Egypt
… 中心 位于 下 埃及 center located-at down Egypt
… 以及 所谓 的 下 埃及 and so-called of down Egypt
… 叫做 下 埃及 called down Egypt

Table 1: Chinese text preceding *Lower Egypt*

Unfortunately, this heuristic does not hold as often as one might imagine. Consider the candidates for *Channel Spacing* in Table 2. The suffix间隔 (gap) has the highest frequency count. It is nonetheless an incomplete translation of *Channel Spacing*. The correct translations in rows *c* to *h* occurred with *Channel Spacing* only once.

a	… □ 为 频道 间距 λ is channel distance
b	… 其 频道 间距 its channel distance
c	… 除了 降低 波道 间距 in-addition-to reducing wave-passage distance
d	… 亦 展示 具 波道 间隔 also showed have wave-passage gap
e	… 也 就 是 频道 间隔 also therefore is channel gap
f	… 且 频道 的 间隔 and channel 's gap
g	… 一个 重要 特性 是 信道 间隔 an important property is signal-passage gap
h	… 已经 能够 达到 通道 间隔 already able reach passage gap

Table 2: Text preceding *Channel Spacing*

The crucial observation we make here is that although the words like 信道 (in row *g*) co-occurred with *Channel Spacing* only once, there are many co-occurrences of 信道and *Channel* in other candidate pairs, such as:

… 而 不 是 语音 信道 (Speech Channel)
… 块 平坦 衰落 信道 (Block Flat Fading Channel)
… 信道 B (Channel B)
… 光纤 信道 探针 (Fiber Channel Probes)

… 反向 信道 (Reverse Channel)
… 基带 滤波 反向 信道 (Reverse Channel)

Unlike previous approaches that rely solely on the preceding text of a single English term to determine its translation, we treat the entire collection of candidate pairs as a partially parallel corpus and establish the correspondences between the words using a word alignment algorithm.

At first glance, word alignment appears to be a more difficult problem than the extraction of parenthetical translations. Extraction of parenthetical translations need only determine the *first* pre-parenthesis word aligned with an in-parenthesis word, whereas word alignment requires the respective linking of *all* such (pre,in)-parenthesis word pairs. However, by casting the problem as word alignment, we are able to generalize across instances involving different in-parenthesis terms, giving us a larger number of, and more varied, example contexts per word.

For the examples in Table 2, the words频道 (channel), 波道(wave passage), 信道(signal passage), and 通道 (passage) are aligned with *Channel*, and the words间距(distance) and 间隔 (gap) are aligned with *Spacing*. Given these alignments, the left boundary of the translated Chinese term is simply the leftmost word that is linked to one of the English words.

Our algorithm consists of two steps:

Step 1 constructs a partially parallel corpus. This step takes as input a large collection of Chinese web pages and converts the sentences with parentheses containing English text into pairs of candidates.

Step 2 uses an unsupervised algorithm to align English and Chinese and identify the term being translated according to the left-most aligned Chinese word. If no word alignments can be established, the pair is not considered a translation.

The next two sections present the details of each of the two steps.

3 Constructing a Partially Parallel Corpus

3.1 Filtering out non-translations

The first step of our algorithm is to extract parentheticals and then filter out those that are not translations. This filtering is required as parenthetical translations represent only a small fraction of the usages for parentheses (see Sec. 5.1). Table 3 shows some example of parentheses that are not translations.

The input to Step 1 is a collection of arbitrary web documents. We used the following criteria to identify candidate pairs:

- The pre-parenthesis text (T_p) is predominantly in Chinese and the in-parenthesis text (T_i) is predominantly in English.
- The concatenation of the digits in T_p must be identical to the concatenation of the digits in T_i. For example, rows *a, b* and *c* in Table 3 can be ruled out this way.
- If T_p contains some text in English, the same text must also appear in T_i. This filters out row *d*.
- Remove the pairs where T_i is part of anchor text. This rule is often applied to instances like row *e* where the file type tends to be inside a clickable link to a media file.
- The punctuation characters in T_p must also appear in T_i, unless they are quotation marks. The example in row *f* is ruled out because '/' is not found in the pre-parenthesis text.

	Examples with translations in *italic*	**Function of the in-parenthesis text**
a	其数值通常在1.4~3.0之间 (MacArthur, 1967) *The range of its values is within 1.4~3.0 (MacArthur, 1967)*	to provide citation
b	越航北京/胡志明 (VN901 15:20-22:30) *Vietnam Airlines Beijing/Ho Chi Minh (VN901 15:20-22:30)*	flight information
c	銷售台球桌（255-8FT） *sale of pool table (255-8FT)*	product Id.
d	// 主程序 // void main (void) *// main program // void main (void)*	function declaration
e	电影名称: 千年湖 (DVD) *movie title: Thousand Year Lake (DVD)*	DVD is the file type
f	水样 所 消耗 的 质量 (g/L) *mass consumed by water sample (g/L)*	measurement unit
g	柔和保养面油 (Sensitive) *gentle protective facial cream (Sensitive)*	to indicate the type of the cream
h	美国九大搜索引擎评测第四章 (Ask Jeeves) *Evaluation of Nine Main Search Engines in the US: Chapter 4 (Ask Jeeves)*	Chapter 4 is about Ask Jeeves

Table 3: Other uses of parentheses

The instances in rows *g* and *h* cannot be eliminated by these simple rules, and are filtered only later, as we fail to discover a convincing word alignment.

3.2 Constraining term boundaries

Similar to (Cao *et al.* 2007), we segmented the pre-parenthesis Chinese text and restrict the term boundary to be one of the segmentation boundaries. Since parenthetical translations are mostly translation of terms, it makes sense to further constrain the left boundary of the Chinese side to be a term boundary. Determining what should be counted as a term is a difficult task and there are not yet well-accepted solutions (Sag *et al,* 2003).

We compiled an approximate term vocabulary by taking the top 5 million most frequent Chinese queries as according to a fully anonymized collection of search engine query logs.

Given a Chinese sentence, we first identify all (possibly overlapping) sequences of words in the sentence that match one of the top-5M queries. A matching sequence is called a maximal match if it is not properly contained in another matching sequence. We then define the **potential boundary positions** to be the boundaries of maximal matches or words that are not covered by any of the top-5M queries.

3.3 Length-based trimming

If there are numerous Chinese words preceding a pair of parentheses containing two English words, it is very unlikely for all but the right-most few Chinese words to be part of the translation of the English words. Including extremely long sequences as potential candidates introduces significantly more noise and makes word alignment harder than necessary. We therefore trimmed the pre-parenthesis text with a length-based constraint. The cut-off point is the first (counting from right to left) potential boundary position (see Sec. 3.2) such that $C \geq 2\ E + K$, where C is the length of the Chinese text, E is the length of the English text in the parentheses and K is a constant (we used K=6 in our experiments). The lengths C and E are measured in bytes, except when the English text is an abbreviation (in that case, E is multiplied by 5).

4 Word Alignment

Word alignment is a well-studied topic in Machine Translation with many algorithms having been proposed (Brown et al, 1993; Och and Ney 2003). We used a modified version of one of the simplest word alignment algorithms called Competitive Linking (Melamed, 2000). The algorithm assumes that there is a score associated with each pair of words in a bi-text. It sorts the word pairs in descending order of their scores, selecting pairs based on the resultant order. A pair of words is linked if none of the two words were previously linked to any other words. The algorithm terminates when there are no more links to make.

Tiedemann (2004) compared a variety of alignment algorithms and found Competitive Linking to have one of the highest precision scores. A disadvantage of Competitive Linking, however, is that the alignments are restricted word-to-word alignments, which implies that multi-word expressions can only be partially linked at best.

4.1 Dealing with multi-word alignment

We made a small change to Competitive Linking to allow consecutive sequence of words on one side to be linked to the same word on the other side. Specifically, instead of requiring both e_i and f_j to have no previous linkages, we only require that at least one of them be unlinked and that (suppose e_i is unlinked and f_j is linked to e_k) none of the words between e_i and e_k be linked to any word other than f_j.

4.2 Link scoring

We used φ^2 (Gale and Church, 1991) as the link score in the modified competitive linking algorithm, although there are many other possible choices for the link scores, such as χ^2 (Zhang, S. Vogel. 2005), log-likelihood ratio (Dunning, 1993) and discriminatively trained weights (Taskar *et al*, 2005). The φ^2 statistics for a pair of words e_i and f_j is computed as

$$\varphi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

where

a is the number of sentence pairs containing both e_i and f_j;

$a+b$ is the number of sentence pairs containing e_i;

$a+c$ is the number of sentence pairs containing f_j;

d is the number of sentence pairs containing neither e_i nor f_j.

The φ^2 score ranges from 0 to 1. We set a threshold at 0.001, below which the φ^2 scores are treated as 0.

4.3 Bias in the partially parallel corpus

Since only the last few Chinese words in a candidate pair are expected to be translated, there should be a preference for linking the words towards the end of the Chinese text. One advantage of Competitive Linking is that it is quite easy to introduce such preferences into the algorithm, by using the word positions to break ties of the φ^2 scores when sorting the word pairs.

4.4 Capturing syllable-level regularities

Many of the parenthetical translations involve proper names, which are often transliterated according to the sound. Word alignment algorithms have generally ignored syllable-level regularities in transliterated terms. Consider again the *Shapiro* example in the introduction section. There are numerous correct transliterations for the same English word, some of which are not very frequent. For example, the word 夏布洛happens to have a similar φ^2 score with *Shapiro* as the word 流利 (fluency), which is totally unrelated to *Shapiro* but happened to have the same co-occurrence statistics in the (partially) parallel corpus.

Previous approaches to parenthetical translations relied on specialized algorithms to deal with transliterations (Cao *et al*, 2007; Jiang *et al*, 2007; Wu and Chang, 2007). They convert Chinese words into their phonetic representations (Pinyin) and use the known transliterations in a bilingual dictionary to train a transliteration model.

We adopted a simpler approach that does not require any additional resources such as pronunciation dictionaries and bilingual dictionaries. In addition to computing the φ^2 scores between words, we also compute the φ^2 scores of prefixes and suffixes of Chinese and English words. For both languages, the prefix of a word is defined as the first three bytes of the word and the suffix is defined as the last three bytes. Since we used UTF-8 encoding, the first and last three bytes of a Chinese word, except in very rare cases, correspond to the first and last Chinese character of the word. Table 4 lists the English prefixes and suffixes that have the highest φ^2 scores with the Chinese prefix 夏and suffix洛.

Type	Chinese	English
prefix	夏	sha, amo, cha, sum, haw, lav, lun, xia, xal, hnl, shy, eve, she, cfh, …
suffix	洛	rlo, llo, ouh, low, ilo, owe, lol, lor, zlo, klo, gue, ude, vir, row, oro, olo, aro, ulo, ero, iro, rro, loh, lok, …

Table 4: Example prefixes and suffixes with top φ^2

In our modified version of the competitive linking algorithm, the link score of a pair of words is the sum of the φ^2 scores of the words themselves, their prefixes and their suffixes.

In addition to syllable-level correspondences in transliterations, the φ^2 scores of prefixes and suffixes can also capture correlations in morphologically composed words. For example, the Chinese prefix 三 (three) has a relatively high φ^2 score with the English prefix *tri*. Such scores enable word alignments to be made that may otherwise be missed. Consider the following text snippet:

..... 三 嗪 氟草胺 (triaziflam)

The correct translation for *triaziflam* is三嗪氟草胺. However, the Chinese term is segmented as 三 + 嗪 + 氟草胺. The association between三 (three) and *triaziflam* is very weak because 三is a very frequent word, whereas *triaziflam* is an extremely rare word. With the addition of the φ^2 score between 三and *tri*, we were able to correctly establish the connection between *triaziflam* and 三.

It turns out to be quite effective to assume prefixes and suffixes of words consist of three bytes, despite its apparent simplicity. The benefit of φ^2 scores for prefixes and suffixes is not limited to morphemes that happen to be three bytes long. For example, the English morpheme "du-" corresponds to the Chinese character 二 (two). Although the φ^2 between *du* and二 won't be computed, we do find high φ^2 scores between二 and *due* and between二 and *dua*. The three letter prefixes account for many of the words with the *du-* prefix.

5 Experimental Results

We extracted from Chinese web pages about 1.58 billion unique sentences with parentheses that contain ASCII text. We removed duplicate sentences so that duplications of web documents will not skew the statistics. By applying the filtering algorithm in Sec. 3.1, we constructed a partially paral-

lel corpus with 126,612,447 candidate pairs (46,791,841 unique), which is about 8% of the number of sentences. Using the word alignment algorithm in Sec. 4, we extracted 26,753,972 translation pairs between 13,471,221 unique English terms and 11,577,206 unique Chinese terms.

Parenthetical translations mined from the Web have mostly been evaluated by manual examination of a small sample of results (usually a few hundred entries) or in a Cross Lingual Information Retrieval setup. There does not yet exist a common evaluation data set.

5.1 Evaluation with Wikipedia

Our first evaluation is based on translations in Wikipedia, which contains far more terminology and proper names than bilingual dictionaries. We extracted the titles of Chinese and English Wikipedia articles that are linked to each other and treated them as gold standard translations. There are 79,714 such pairs. We removed the following types of pairs because they are not translations or are not terms:

- Pairs with identical strings. For example, both English and Chinese versions have an entry titled ".ch";
- Pairs where the English term begins with a digit, e.g., "245", "300 BC", "1991 in film";
- Pairs where the English term matches the regular expression 'List of .*', e.g., "List of birds", "List of cinemas in Hong Kong";
- Pairs where the Chinese title does not have any non-ASCII code. For example, the English entry "Syncfusion" is linked to ".NET Framework" in the Chinese Wikipedia.

The resulting data set contains 68,131 translation pairs between 62,581 Chinese terms and 67,613 English terms. Only a small percentage of terms have more than one translation. Whenever there is more than one translation, we randomly pick one as the answer key.

For each Chinese and English word in the Wikipedia data, we first find whether there is a translation for the word in the extracted translation pairs. The **Coverage** of the Wikipedia data is measured by the percentage of words for which one or more translations are found. We then see whether our most frequent translation is an **Exact Match** of the answer key in the Wikipedia data.

	Coverage	Exact Match
Full	**70.8%**	**36.4%**
-term	67.1%	34.8%
-pre-suffix	67.6%	34.4%
IBM	67.6%	31.2%
LDC	10.8%	4.8%

Table 5: Chinese to English Results

	Coverage	Exact Match
Full	**59.6%**	**27.9%**
-term	**59.6%**	27.5%
-pre-suffix	58.9%	27.4%
IBM	52.4%	13.4%
LDC	3.0%	1.4%

Table 6: English to Chinese Results

Table 5 and 6 show the Chinese-to-English and English-to-Chinese results for the following systems:

Full refers to our system described in Sec. 3 and 4;

-term is the system without the use of query logs to restrict potential term boundary positions (Sec. 3.2);

-pre-suffix is the system without using the φ^2 score of the prefixes and suffixes;

IBM refers to a system where we substitute our word alignment algorithm with IBM Model 1 and Model 2 followed by the HMM alignment (Och and Ney 2003), which is a common configuration for the word alignment components in machine translations systems;

LDC refers to the LDC2.0 English to Chinese bilingual dictionary with 161,117 translation pairs.

It can be seen that the use of queries to constrain boundary positions and the addition of φ^2 scores of prefixes/suffixes improve the percentage of Exact Match. The IBM Model tends to make many more alignments than Completive Linking. While this is often beneficial for machine translation systems, it is not very suitable for creating bilingual dictionaries, where precision is of paramount importance. The LDC dictionary was manually compiled from diverse resources within LDC and (mostly) from the Internet. Its coverage of Wikipedia data is extremely low, compared to our method.

English	Wikipedia Translation	Parenthetical Translation
Pumping lemma	泵引理	引理[1]
Topic-prominent language	话题优先语言	突出性语言[1]
Yoido Full Gospel Church	汝矣岛纯福音教会	全备福音教会[1]
First Bulgarian Empire	第一保加利亚帝国	强大的保加利亚帝国[2]
Vespid	黄蜂	针对境内胡蜂[2]
Ibrahim Rugova	易卜拉欣·鲁戈瓦	鲁戈瓦[3]
Jerry West	杰里·韦斯特	威斯特[3]
Nicky Butt	尼基·巴特	巴特[3]
Benito Mussolini	贝尼托·墨索里尼	墨索里尼[3]
Ecology of Hong Kong	香港生态	本文介绍的[*]
Paracetamol	对乙酰氨基酚	扑热息痛[*]
Thermidor	热月	必杀[*]
Udo	独活	乌多
Public opinion	舆论	公众舆论
Michael Bay	麦可·贝	迈克尔·贝
Dagestan	达吉斯坦共和国	达吉斯坦
Battle of Leyte Gulf	莱特湾海战	莱伊特海湾战役
Glock	格洛克手枪	格洛克
Ergonomics	人因工程学	工效学
Frank Sinatra	法兰·仙纳杜拉	法兰克辛纳屈
Zaragoza	萨拉戈萨省	萨拉戈萨
Komodo	科莫多岛	科摩多岛
Eli Vance	伊莱·万斯	伊莱·凡斯博士
Manitoba	缅尼托巴	曼尼托巴省
Giant Bottlenose Whale	阿氏贝喙鲸	巨瓶鼻鲸
Exclusionary rule	证据排除法则	证据排除规则
Computer worm	蠕虫病毒	计算机蠕虫
Social network	社会性网络	社会网络
Glasgow School of Art	格拉斯哥艺术学校	格拉斯哥艺术学院
Dee Hock	狄伊·哈克	迪伊·霍克
Bondage	绑缚	束缚
The China Post	英文中国邮报	中国邮报
Rachel	拉结	瑞秋
John Nash	约翰·纳西	约翰·纳什
Hattusa	哈图沙	哈图萨
Bangladesh	孟加拉国	孟加拉

Table 7: A random sample of non-exact-matches

[1]the extracted translation is too short
[2]the extracted translation is too long
[3]the extracted translation contains only the last name
[*]the extracted term is completely wrong.

Note that Exact Match is a rather stringent criterion. Table 7 shows a random sample of extracted parenthetical translations that failed the Exact Match test. Only a small percentage of them are genuine errors. We nonetheless adopted this measure because it has the advantage of automated evaluation and our goal is mainly to compare the relative performances.

To determine the upper bound of the coverage of our web data, for each Wikipedia English term we searched within the total set of available parenthesized text fragments (our English candidate set before filtering as by Step 1). We discovered 81% of the Wikipedia titles, which is approximately 10% above the coverage of our final output. This indicates a minor loss of recall because of mistakes made in filtering (Sec. 3.1) and/or word alignment.

5.2 Evaluation with term translation requests

To evaluate the coverage of output produced by their method, Cao *et al* (2007) extracted English queries from the query log of a Chinese search engine. They assume that the reason why users typed the English queries in a Chinese search box is mostly to find out their Chinese translations. Examining our own Chinese query logs, however, the most-frequent English queries appear to be navigational queries instead of translation requests. We therefore used the following regular expression to identify queries that are unambiguously translation requests:

```
/^[a-zA-Z ]* 的中文$/
```

where的中文means "*'s Chinese*". This regular expression matched 1579 unique queries in the logs. We manually judged the translation for 200 of them. A small random sample of the 200 is shown in Table 8. The empty cells indicate that the English term is missing from our translation pairs. We use * to mark incorrect translations. When compared with the sample queries in (Cao *et al.*, 2007), the queries in our sample seem to contain more phrasal words and technical terminology. It is interesting to see that even though parenthetical translations tend to be out-of-vocabulary words, as we have remarked in the introduction, the sheer size of the web means that occasionally translations of common words such as 'use' are sometimes included as well.

buckingham palace	白金汉宫
chinadaily	中国日报
coo	首席运营官
diammonium sulfate	
emilio pucci	埃米里奥·普奇
finishing school	精修学校
gloria	格洛丽亚
horny	长角收割者*
jam	詹姆
lean six sigma	精益六西格玛
meiosis	减数分裂
near miss	迹近错失
pachycephalosaurus	肿头龙
pops	持久性有机污染物
recreation vehicle	休闲露营车
shanghai ethylene cracker complex	
stenonychosaurus	细爪龙
theanine	茶氨酸
use	使用
with you all the time	回想和你在一起的日子里

Table 8: A small sample of manually judged query translations

We compared our results with translations obtained from Google and Yahoo's translation services. The numbers of correct translations for the random sample of 200 queries are as follows:

Systems	**Google**	**Yahoo!**	**Mined**	**Mined+G**
Correct	115	84	116	135

Our system's outputs (**Mined**) have the same accuracy as the Google Translate. Our outputs have results for 154 out of the 200 queries. The 46 missing results are considered incorrect. If we combine our results with Google Translate by looking up Google results for missing entries, the accuracy increases from 56% to 68% (**Mined+G**). If we treat the LDC Chinese-English Dictionary 2.0 as a translator, it only covers 20.5% of the 200 queries.

5.3 Evaluation with SMT

The extracted translations may serve as training data for statistical machine translation systems. To evaluate their effectiveness for this purpose, we trained a baseline phrase-based SMT system (Koehn *et al*, 2003; Brants *et al*, 2007) with the FBIS Chinese-English parallel text (NIST, 2003). We then added the extracted translation pairs as additional parallel training corpus. This resulted in a 0.57 increase of BLEU score based on the test data in the 2006 NIST MT Evaluation Workshop.

6 Related Work

Nagata *et al.* (2001) made the first proposal to mine translations from the web. Their work was concentrated on terminologies, and assumed the English terms were given as input. Wu and Chang (2007), Kwok *et al.* (2005) also employed search engines and assumed the English term given as input, but their focus was on name transliteration. It is difficult to build a truly large-scale translation lexicon this way because the English terms themselves may be hard to come by.

Cao *et al.* (2007), like us, used a 300GB collection of web documents as input. They used supervised learning to build models that deal with phonetic transliterations and semantic translations separately. Our work relies on unsupervised learning and does not make a distinction between translations and transliterations. Furthermore, we are able to extract two orders of magnitude more translations from than (Cao *et al.*, 2007).

7 Conclusion

We presented a method to apply a word alignment algorithm on a partially parallel corpus to extract translation pairs from the web. Treating the translation extraction problem as a word alignment problem allowed us to generalize across instances involving different in-parenthesis terms. Our algorithm extends Competitive Linking to deal with multi-word alignments and takes advantage of word-internal correspondences between transliterated words or morphologically composed words. Finally, through our discussion of parallel Wikipedia topic titles as a gold standard, we presented the first evaluation of such an extraction system that went beyond manual judgments on small sized samples.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments.

References

T. Brants, A. Popat, P. Xu, F. Och and J. Dean, *Large Language Models for Machine Translation*, *EMNLP-CoNLL*-2007.

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. *The mathematics of statistical machine translation: Parameter estimation*. *Computational Linguistics,* 19(2):263–311.

G. Cao, J. Gao and J.Y. Nie. 2007. *A system to mine large-scale bilingual dictionaries from monolingual Web pages, MT Summit*, pp. 57-64.

T. Dunning. 1993. *Accurate Methods for the Statistics of Surprise and Coincidence*. *Computational Linguistics* 19, 1.

W. Gale and K. Church. 1991. *Identifying word correspondence in parallel text*. In *Proceedings of the DARPA NLP Workshop*.

L. Jiang, M. Zhou, L.F. Chien, C. Niu. 2007. *Named Entity Translation with Web Mining and Transliteration.* In *Proc. of IJCAI-2007*. pp. 1629-1634.

P. Koehn, F. Och and D. Marcu, *Statistical Phrase-based Translation*, In *Proc. of HLT-NAACL 2003*.

K.L. Kwok, P. Deng, N. Dinstl, H.L. Sun, W. Xu, P. Peng, and J. Doyon. 2005. *CHINET: a Chinese name finder system for document triage.* In *Proceedings of 2005 International Conference on Intelligence Analysis*.

I.D. Melamed. 2000. *Models of translational equivalence among words*. *Computational Linguistics*, 26(2):221–249.

M. Nagata, T. Saito, and K. Suzuki. 2001. *Using the Web as a bilingual dictionary*. In *Proc. of ACL 2001 DD-MT Workshop*, pp.95-102.

NIST. 2003. The NIST machine translation evaluations. http://www.nist.gov/speech/tests/mt/.

F.J. Och and H. Ney. 2003. *A systematic comparison of various statistical alignment models. Computational Linguistics,* 29(1):19–51.

I.A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. *Multiword expressions: A pain in the neck for NLP*. In *Proc. of CICLing-2002*, pp 1–15, Mexico City, Mexico.

B. Taskar, S. Lacoste-Julien, and D. Klein. 2005. *A discriminative matching approach to word alignment*. In *Proc. of HLT/EMNLP-05*. Vancouver, BC.

J. Tiedemann. 2004. *Word to word alignment strategies*. In *Proceedings of the 20th international Conference on Computational Linguistics*. Geneva, Switzerland.

J.C. Wu and J.S. Chang. 2007. *Learning to Find English to Chinese Transliterations on the Web*. In Proc. of EMNLP-CoNLL-2007. pp.996-1004. Prague, Czech Republic.

Y. Zhang, S. Vogel. 2005 *Competitive Grouping in Integrated Phrase Segmentation and Alignment Model*. in Proceedings of *ACL-05 Workshop on Building and Parallel Text*. Ann Arbor, MI.