Extracting Bilingual Terms from Mainly Monolingual Data

Francis Bond　Zhiqiang Chang*　**Kiyotaka Uchimoto**
National Insititute of Information and Communications Technology
* Kyoto University
bond@ieee.org, myfuture0416@hotmail.co.jp, uchimoto@nict.go.jp

1 Introduction

In this paper we present a way to extract bilingual terms from mainly mono-lingual data that exploits explicit in-text cues, in this case the use of parentheses "()" and character type. We extract Japanese-English terms from Japanese text, and Chinese-English terms from Chinese text. We show the results of extracting terms from various genres, including web, newspaper and academic text.

As access to information in foreign languages has increased, so too, has the emand for translation and multilingual lexical resources, such as dictionaries. It is difficult to keep expanding lexicons, especially dealing with neoligisms in other languages. The increase in cooperative construction of lexicons, such as JMDict (Breen, 2004), Yakushite-net (Sukehiro et al., 2001) and Wikipedia/Wiktionary go some way to solving this problem, but multi-lingual resources still have incomplete cover.

One successful approach is to extract translations from bilingual data, either aligned (Fung, 1995) or comparable(Tanaka and Iwasaki, 1996). However, there is not enough bilingual text to find translations for all possible terms. Nagata et al. (2001) pointed out that, for some language combinations, it is possible to get translations from mainly monolingual data (or, as he put it, partially bilingual data). For example, in Japanese, when introducing a term with a well known English equivalent, such as a technical term or name, it is common to add the English equivalent in brackets after the first usage. For example, in (1), adapted from a Japanese paper on computational linguistics, the term 生成的辞書 "seiseiteki jisho" is explicitly glossed in English as "generative lexicon". In this paper, we call the word being translated the **term**, and its translation the **gloss**.

(1) Pustejovsky の 生成的 辞書 (*generative*
Pustejvsky of seiseiteki jisho (generative
Pustejvsky of generative dictionary (generative
lexicon) の 記述方式 を 利用して
lexicon) no kijutsuhōshiki wo riyōshite
lexicon) of way of representing ACC using
"Using the representation of Pustejovsky's generative lexicon [...]"

The approach of Nagata et al. (2001) was to look at all English words in the text containing the Japanese term you wish to translate, and look for the closest. This had a fairly low accuracy of 18% for the best candidate. Li et al. (2003) use a similar approach.

Our approach is to raise the precision by using more explicit cues (a word and its gloss **in brackets**) and take the search off-line: we will look for all occurrences of "word (gloss)" and compile them into a lexicon. There are several advantages to taking this extraction off-line: (1) not all data is on-line, in particular there are many collections of academic text than cannot be freely accessed online; (2) we can get more reliable frequency data; (3) the terms are available when we want them, there is no need to look up words on the fly; and (4) we can learn patterns over all term-gloss pairs. The main drawback is that the world wide web is continuously updated, and thus may have terms that on offline repository does not have. We are alleviating this by using a large web corpus, which is periodically updated.

2 Basic Approach

Because we are interested in finding bilingual equivalences, we also use the character type. We also use the text direction (we assume the gloss comes after the term), although it would be simple to parametrize this to deal with languages written from right to left. Finally, we assume that the gloss will be in English. This is not always the case (we found examples in Mongolian, German, French and pinyin) as we discuss below. If we

were interested in monolingual relations, then we could also look at similar character types and pick up synonym pairs such as "National Institute for Telecommunications Technologies (NICT)".

We extract terms using two patterns (a) fully bracketed, as in (2) and (b) partly bracketed, as in (3).[1]

(2) Fully Bracketed Examples

a. 「収穫逓減の法則(*the law of diminishing return*)」 (ja)

b. 《德拉吉报道》(*DrudgeReport*) (zh)

c. "魔兽世界"(*World of Warcraft*) (zh)

(3) Partly Bracketed Examples

a. 図 1 に，明瞭性 (*Clarity*)・新奇性 (*Novelty*) (ja)

b. 目标递归策略 （GoalRecursionStrategy），这是一种内部指导的策略。 (zh)

Fully bracketed examples explicitly give both the term and its gloss, but are less frequent than partially bracketed terms. For partially bracketed terms the left-hand limit of the term also has to be determined. In both cases, there is also the problem of deciding the true relation between the term and the gloss.

2.1 Details of matching

We preprocess all input by converting into unicode and splitting into sentences.

We then match the following regular expressions. Basically we accept anything except punctuation for the term, and three or more latin letters (half or full width) along with connector punctuation and whitespace for the gloss.[2]

```
term  =  any non punctuation
gloss = latin, connector punctuation,
        full space latin, whitespace
lbr   = ( （
rbr   = ) ）
tlbr  = Unicode: Punctuation, Open
trbr  = Unicode: Punctuation, Close

full1 = tlbr(term+)trbr lbr(gloss{3,})rbr
full2 = tlbr(term+)lbr(gloss{3,})rbr trbr
part  = (term+)lbr(gloss{3,})rbr
```

The terms are then filtered so that the term must contain at least one CJK character (Chinese character, Hiragana or Katakana) and the gloss must not be in the following stop lists:

Roman Numerals: xii, iii, ...
Units: MPa, Kmh, ...
Smilies: T_T, _o_, m_m, x_x ...
Week Days: mon, wed, fri, ...
Other: pdf, PDF ...

We then run the regular expressions over all sentences, and store the matches as quintuples of ⟨`term, gloss, matchtype, sentence, file`⟩. The `matchtype` is either full or part, the `sentence` is the whole sentence the match applied to, `file` is a string identifying the file and position in which the sentence appears.

2.2 Data Availability

The ⟨term, gloss, matchtype⟩ triples can be downloaded from from `/www2.nict.go.jp/x/x161/en/member/bond/data`.

3 Experiments

We extracted bilingual terms from the following corpora (summarized in Table 1). For Japanese we used a corpus of papers from the Journal of Natural Language Processing (NLP), a corpus of papers from the Japan Science and Technology Agency (J-STAGE: `www.jstage.jst.go.jp/browse/-char/ja`), and an enlarged version of the WEB corpus used in Kawahara and Kurohashi (2006) (WWW). For Chinese we used the Beijing Language and Culture University corpus of mainly journal articles (BLCU) and the SohuNews corpus of IT news (Sohu).

The raw numbers of term gloss pair instances found are given in Table 2. The number of fully

Lang	Name	Size (MB)
Ja	WWW	514,212
	J-STAGE	604
	NLP	43
Zh	BLCU	80,000
	Sohu	974

Table 1: Size and types of Corpora Used

[1] Examples are shown with the language being extracted: zh = Chinese, ja = Japanese.

[2] In addition to Japanese and Chinese, we have also successfully tested these patterns on Thai.

Lang	Name	Full	Part
Ja	WWW	896,000	14,861,000
	J-STAGE	552	45,000
	NLP	64	1,300
Zh	BLCU	151,000	6,563,000
	SohuTechNews	5,400	33,000

Table 2: Distribution of Bracketed Terms

#	English	Chinese
10	World of Warcraft	魔兽世界
5	WOW	魔兽世界
3	WoW	魔兽世界
2	WorldofWarcraft	魔兽世界
1	World orWarcraft	魔兽世界
1	World of WarcraftTM	魔兽世界
1	Warcraft	魔兽世界

Table 3: Fully Bracketed Terms from Sohu

bracketed terms is an order of magnitude (6-80) times less than the number of partially bracketed terms.

4 Results

In this section we examine the results in more detail.

4.1 Fully Bracketed

Examining the extracted fully bracketed terms showed that most of them were potentially good pairs, but that there was considerable noise in the corpus. For example, consider the terms extracted for 魔兽世界 *MuoShouShiJie* "World of Warcraft" given in Table 3. The most frequent three candidates are all possible translation, but the remainder are problematic, with whitespace missing, spelling errors and so on.

In a case were there are many instances, we can choose the most frequent, but this not possible for low frequency terms.

There was variation in case (*atom* vs *Atom*), number (*atom* vs *atoms*), absence or presence of white space, absence or presence of articles (*the atom* vs *atom*), different derivational forms (*atom* vs *atomic*) and plain spelling errors. We hypothesize that the major source of the errors is the use of OCRs in creating the corpora. Because the primary language of the texts is not English, the OCR does worse with the English strings. In particular, because neither Japanese or Chinese segment words, the OCR fails to separate the English words.

Currently we merge similar entries, combining entries that differ only in case, white space, the presence or absence of articles and number.

In future work, we plan to also merge acronyms (*WoW* vs ***W****orld* ***o****f* ***W****arcraft*, derivational forms and minor spelling errors.

There were also several examples where the gloss was not English, for example 主体, where we had both *subject* (with variants *Subject* and *subjects* and German *subjekt* with variant (*Subjekt*), however, they were few enough to be ignored.

The main remaining problem is the lack of spaces, we are currently investigating methods of adding spaces back in.

4.2 Partially Bracketed

For the partially bracketed results, as well as the problems encountered with the glosses above, there is also the problems associated with finding the left boundary of the term.

Consider the strings extracted from (1): "Pustejovskyの生成的辞書" and "*generative lexicon*". The modifying phrase *Pustejovsky*の "Pustejovsky's" is not part of the term, but is matched by the regular expression. We can solve this by two methods. First, given enough examples, we can discard non-shared left hand contexts, as in (4). Only the shared string is part of the term. Second, we can look for left hand contexts that generally aren't part of a term, such as the accusative marker を *wo*, and discard them and everything to their left. We use these strategies to merge similar entries.

(4) a. ~~Pustejovsky の~~生成的辞書
b. ~~*Telic*とは~~生成的辞書

4.3 Merged Results

The number of terms after merging is shown in Table 4. We have extracted over a million terms in each language.

Examples from the BLCU and JST corpora, sampled at intervals of 1,000, ranked in descending order of frequency, are given in Table 5. Examples of bad extracted pairs are marked with an asterisk. There are no bad examples in the high frequency set from BLCU. The two bad examples

Lang	Name	# Merged
Ja	WWW	1,635,000
	J-STAGE	20,000
	NLP	372
Zh	BLCU	964,000
	Sohu	33,000

Table 4: Results after Merging

Rank	English	Ja/Zh	Freq
BLCU			
1	*SOD*	超氧化物歧化	18,000
1001	*quercetin*	槲皮素	121
2001	*Alcan*	加拿大公司	55
3001	*CSTC*	中国件中心	34
4001	*John*	约翰	18
5001	*Username*	户名	18
JST			
1	*Bunseki Kagaku*	分析化学	517
1001	*STEM*	走査型TEM	2
2001	*structural factor*	構造係数	2
3001	*explicit attitude*	*的態度	1
4001	*Lake Magadi*	マガディ湖	1
5001	*ALPase*	*~~で培養細胞の~~ アルカリホスファターゼ	1

Table 5: Extracted Examples after Merging

from the JST corpus come from a mistake by the sentence splitter, and an inability to find the correct left hand boundary.

For the final dictionary, we combine all the raw results before the merging. This gives more examples to use in determining the best terms.

5 Evaluation

We did an detailed evaluation on the results from the NLP corpus. Each term was checked by two people (one author and one non-author). We divided the terms into four classes:

Known good terms already in our lexicons[3]
文法機能 *grammatical function*

New good terms not in any of our lexicons[4]
生成的辞書 *generative lexicon*

General good translations but not NLP terms
すべての学生 *all of the students* (Example)

Other the remainder
似テイル *ohxap ketidu* (Mongolian)
形態素解析システム *JUMAN* (Description)

[3]EDR, JMDict, CICC and lingdic.

[4]These were then added to lingdic `www2.nict.go.jp/x/x161/en/member/bond/data/lingdic`.

Status	#	%
Known	61	16%
New	138	37%
General	74	20%
Other	99	27%
Total	372	100%

Table 6: Results for the NLP Corpus

37% were the most useful results: new terms not in our lexicons. Another 20% were good translations, even if not NLP terms. Only 27% were not useful, and even then many could be made useful with more processing.

6 Conclusions and Future Work

We have used simple robust in-text cues to create a large bilingual dictionary from mainly monolingual data. In future work, we would like to (a) run our extraction method over more corpora in more languages, (b) improve the refinement method: in particular look at ways to judge candidates with a frequency of one. Currently we are investigating features such as compositionality, length, and transliterated equivalence.

References

J. W. Breen. 2004. JMDict: a Japanese-multilingual dictionary. In *Coling 2004 Workshop on Multilingual Linguistic Resources*, pages 71–78. Geneva.

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *33th Annual Conference of the Association for Computational Linguistics*, pages 236–243.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 67–73.

Hang Li, Yunbo Cao, and Cong Li. 2003. Using bilingual web data to mine and rank translations. *IEEE Intelligent Systems*, 18(4):54–59.

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the web as a bilingual dictionary. In *ACL Workshop on Data-driven Methods in Machine Translation*, pages 95–102.

Tatsuya Sukehiro, Mihoko Kitamura, and Toshiki Murata. 2001. Collaborative translation environment 'Yakushite.Net'. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium: NLPRS-2001*, pages 769–770. Tokyo.

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *16th International Conference on Computational Linguistics: COLING-96*, pages 580–585. Copenhagen.