Towards Bilingual Term Extraction in Comparable Patents *

Bin Lu and Benjamin K. Tsou

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
lubin2@student.cityu.edu.hk, rlbtsou@cityu.edu.hk

Abstract. In order to extract bilingual terms in a corpus of comparable patents, we present a novel framework in this paper. The framework includes the following major steps: 1) extract monolingual single-word and multi-word term candidates in monolingual patents; 2) Find parallel sentences in comparable patents; 3) extract bilingual single-word and multi-word term candidates; 4) identify correct bilingual terms using a SVM classifier by integrating both linguistic and statistical information. The experimental results show that the framework can well identify correct bilingual terms from comparable patents, and the SVM classifier can further improve its performance.

Keywords: Bilingual Term Extraction, Comparable patents, Multi-word Term

1 Introduction

Bilingual term extraction is to extract parallel technical terms from bilingual domain-specific corpora, and it is crucial for many NLP fields, such as Machine Translation, bilingual lexicography, cross-language information retrieval, and bilingual ontology extraction. Many researchers have done bilingual term extraction, such as Kupiec (1994), Daille et al. (2004), Wu and Xia (1994), Vintar(2001), Piperidis and Harlas (2006), Ha et al. (2008).

Previous studies focused on mining bilingual terms from a bilingual sentence-level parallel corpus. However, obtaining a large-scale bilingual sentence-level parallel corpus is very expensive while it is easy to obtain a bilingual comparable corpus. This paper aims to mine bilingual terms from a bilingual comparable corpus in patent domain, and present a novel framework to extract bilingual terms in comparable patents. Patents, which contain a large amount of technical terms, could be one of the major sources for term extraction. In particular, multilingual patents could be used to extract parallel technical terms. The framework for bilingual term extraction includes the following steps:

1) Extract monolingual single-word and multi-word term candidates in monolingual patents by identifying certain English noun phrases, Chinese noun phrase and Chinese verb phrases;

2) Find parallel sentences in comparable patents;

3) Extract bilingual single-word and multi-word term candidates;

4) Identify correct bilingual terms via machine learning approaches by integrating both linguistic and statistical information, including POS-tags and translation probability.

Experiment results indicate that the framework can well identify correct bilingual terms in comparable patents, including both single-word terms and multi-word terms. At the same time, the linguistic and statistic features can be used to further improve the performance of bilingual term extraction via the machine learning approach (e.g. an SVM classifier in this paper).

In the following part, we first describe the Chinese-English comparable patents in Sec. 2, monolingual term candidate generation in Sec. 3, and sentence alignment and filtering in Sec. 4.

* We wish to thank our colleagues Mr. Tao Jiang, Mr. Kapo Chow and Mr. Yuwing Chan for their help of processing the data, Matrixware Information Services GmbH in Vienna for the provision of most of the English patents, Dr. Jingbo Zhu of Northeastern University, China for the invaluable comments.

Extraction of bilingual term candidates, including single-word terms and multi-word terms, is introduced in Sec. 5. The SVM classifier for identifying of correct bilingual terms is introduce in Sec. 6. Lastly we discuss and conclude this paper.

2 Chinese-English Comparable Patents and their Preprocessing

We recently gained access to about 7000 Chinese-English comparable patents, and about 6000 of them contain full texts. The comparable patents belong to the same patent family, and were identified by the priority information described in the English patents (Lu et al., 2009). Each patent has different parts, i.e. *title*, *abstract*, *claim*, *description*, etc. we align sentences in each part of comparable patents. The patents were first segmented into sentences according to punctuations, and the detailed statistics for each section are shown in Table 1.

Table 1: Statistics for each section

Sections	#Chinese Sentences	#English Sentences
Title	7K	7K
Abstract	29K	32K
Claim	145K	201K
Description	557K	840K
Total	738K	1,080K

English sentences were tokenized and POS-tagged by the POS-tagger developed at Stanford NLP lab (Toutanova and Manning, 2000; Toutanova et al., 2003). The tagger uses Penn Treebank POS tagset, including the tags for content words: JJ (adjective), NN (noun), VB (verb), RB (adverb). Chinese sentences were segmented into words and POS-tagged by using a Chinese lexical analyzer ICTCLAS[1]. The POS tag set of ICTCLAS contains 22 first-level POS tags, in which the tags for content words include n (noun), v (verb), a (adjective), b (adjectives to describe difference, e.g. 急性(acute) vs 慢性(chronic)), z (adjective to describe status, e.g. 优良(excellent)), and d (adverb). The numbers of word tokens and types are given in Table 2.

Table 2: Word statistics

	#Word Tokens	#Word Types	#Content Word Types	#Pairs of Word and POS
EN	26.8M	84K	41K[2]	96K
CN	25.6M	64K	17K	64K

3 Identification of Monolingual Term Candidates

Monolingual terms could be single-word or multi-word. For single-word terms, we can just use content words as candidates; while for multi-word terms, it is more complicated, and will be introduced in the following.

4 3.1 Monolingual Multi-word Term Candidates

There are many methods to extract monolingual multi-word term candidates from texts. We just consider noun phrases as term candidates, and extract English and Chinese noun phrases from comparable patents by using regular expressions. The number of words within one phrase was limited to five.

English noun phrases are extracted from English patents by using regular expressions of _JJ*NN+_ such as "NN/NN" as in "coverage area" and "JJ/NN/NN" as in "effective power factor". Chinese noun phrases are extracted from Chinese patents by using regular expressions of _(n|v|a|z|b|d)*(n|v)+_. The reason we include _v (verb)_ in the Chinese regular expressions is that: 1) many verbs can be used as nouns in Chinese; 2) verbs are usually tagged as verbs, even if they are used as nouns; 3) we want to improve the coverage of Chinese term candidates since English term candidates are more easier to identify than Chinese ones. The Chinese candidate

[1] http://ictclas.org/

[2] The English words are already stemmed by using Porter Stemmer.

examples include “离子/n 化合物/n”(ion compound), “摆动/v 装置/n” (swing means), “按摩/v 治疗/v” (massage treatment), “氧化/v 隔离/v 结构/n” (oxide isolation structure).

Table 3: Statistics of term candidates

	#N-gram	#Candidates
EN	26.8M	695K
CN	29.1M	2,690K

Since monolingual terms are not our goal in this study, these candidates will not be filtered by some statistic or linguistic measures. We just use them for the extraction of bilingual terms.

5 Finding Sentence Alignment

To extract sentence pair candidates from the comparable corpus, Champollion[3] is chosen as the sentence aligner. For the bilingual dictionary needed by Champollion, we combine LDC_CE_DIC2.0[4] constructed by LDC, bilingual terms in HowNet[5] and the bilingual lexicon in Champollion.

In total, 355K sentence pair candidates are extracted by Champollion. We randomly sampled 1,000 pairs from the candidates and two Chinese-English bilingual annotators were asked to classify them into three categories: *correct*, *partially correct*, and *incorrect.* The final numbers of the three categories are 448 (44.8%), 114 (11.4%) and 438 (43.8%), respectively. The above numbers show that a large proportion of aligned sentences are incorrect because of noise in comparable patents.

4.1 Filtering of Aligned Sentence

To filter out incorrect alignments, we sorted all sentence pairs based on a confidence scoring metrics so as to filter out those with lower ranking as incorrect alignments (see Lu et al. 2009 for more details). We combined three individual measures to sort sentence pairs: 1) the length-based score P_l; 2) the dictionary-based score P_d; 3) the translation probability score P_t.

Suppose we are given a sentence pair, namely the Chinese sentence S_c and its English counterpart S_e, and l_c and l_e respectively denote the lengths of S_c and S_e in terms of the number of words. Three kinds of measures for scoring aligned sentences are introduced as follows.

1) The ***length-based score*** P_l *(Len)*: we consider the length ratio between S_c and S_e has a normal distribution with mean μ and variance σ^2 (Gale and Church, 1991). The parameters μ and σ^2 are estimated on the preliminary sentence pairs obtained in Sec. 4.

2) The ***dictionary-based score*** P_d : the score is used to compute the content similarity of the sentence pair based on a bilingual dictionary (Utiyama and Isahara, 2003). For the Chinese-Engish dictionary, we just use the one mentioned in Sec. 4.1.

3) The ***bidirectional translation probability score*** P_t *(Tran)*: it combines the translation probability value of both directions (i.e. Chinese->English and English->Chinese), instead of using only one direction (Moore, 2002; Chen, 2003). The preliminarily aligned sentences mentioned in Sec. 4 was used as the training data and compute the word alignment probability score given by the default training process of Giza++ (Och and Ney, 2003)

[3] http://champollion.sourceforge.net/

[4] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

[5] http://www.keenage.com/html/e_index.html

To combine the three measures, we just remove some sentence pairs if use P_d or P_l is lower than the corresponding predefined threshold, and use P_t to sort other remaining sentence pairs by descending order. We randomly selected 100 samples from each of the 12 blocks ranked at the top 240,000 sentence pairs (each block has 20,000 pairs). An annotator classified them into *correct (Cor)*, *partially correct (PaC)*, and *incorrect (IC)* just as in Sec. 3. The results of evaluation are given in Table 4.

Table 4: Rank vs judgement

Range	#Cor	#PaC	#IC
1 -	98	1	1
20001 -	98	0	2
40001 -	96	2	2
60001 -	91	5	4
80001 -	92	2	6
100001 -	88	1	11
120001 -	77	6	17
140001 -	73	7	20
160001 -	64	7	29
180001 -	37	7	56
Total	814	38	148

The table shows that the number of *IC's* increases rapidly as the rank increases. This demonstrates that the ensemble method *Filter* can differentiate the correct alignments from the incorrect ones. Since there is a big jump of the number of *IC'*s between 16001-18000 and 18001-20000, we choose the top 160K alignments as the final parallel corpus, in which the average precision of correct and partially correct sentences is about 90.0% based on the samples above.

6 Extraction of Bilingual Term Candidates

With these 160K aligned sentences, we do word alignment with Giza++ (Och and Ney, 2003) which combines IBM models (Brown et al., 1994) and HMM for the bilingual data. The bilingual corpus is aligned bi-directionally, i.e. first we perform word alignment from English to Chinese and second from Chinese to English. Then these two alignments are combined to form the final word alignment with the heuristics used in the Moses toolkit (Koehn et al., 2007).

From the bidirectional word alignment, we extracted phrasal translation pairs that are consistent with the word alignment by using Moses. This means that words in a phrase pair are only aligned to each other, without non-aligned words. The maximum phrase length used in our experiments was set to five. Meantime, we remove as noise those word/phrase pairs whose translation probability at either direction (Chinese->English or English->Chinese) is lower than 0.1.

5.1 Bilingual Single-word Term Candidates

With Moses, we can get a table of lexical translations, in which both sides of a translation pair are single words, not phrases. For these word alignments, we use the combined bilingual dictionary mentioned in Sec. 4 to filter those common words and consider the remaining ones as term candidates. Some bilingual single-word examples are shown in Table 5. The first column shows the English terms; the second Chinese terms; the third the translation probability of the Chinese phrase to the English phrase; and the fourth vice versa. In total, 9093 English single words and their Chinese equivalents are found in the translation table with probability higher than 0.1. We manually verified these single-word terms, and 48.5% of them are correct translations.

Table 5: Bilingual single-word term examples

English Term	Chinese Term	CN->EN Prob.	EN->CN Prob.
Acclimation	驯化	0.33	0.33
accompaniment	伴唱	0.43	0.4
Accordion	手风琴	0.92	0.52
Accountant	会计	0.5	0.36
Acetic	醋酸	0.35	0.32
Acetochloral	三氯乙醛	0.33	0.5
Acetone	丙酮	0.63	0.63

5.2 Bilingual Multi-Word Term Candidates

To extract multi-word term candidates, we now have three lists: 1) the English term candidates; 2) the Chinese term candidates; 3) bilingual phrasal translation pairs got by Moses. First, we extracted the bilingual translation pairs whose English side strings are in the list of English term candidates, and denote these pairs as *EBil*. Then we filter out those pairs whose Chinese side strings are not in the list of Chinese term candidates from *ENBil*, and denote the remaining candidates as *ECBil.*

In total, Moses find about 5,310K bilingual phrasal pairs, out of which 1,236K (23.3%) pairs have greater probability than 0.1 at both directions (CN->EN and EN->CN), and *EBil* comprises 71,621 (5.4% out of 1,236K) pairs. Table 6 provides several samples of extracted bilingual phrasal pairs. The first column shows the English terms; the second Chinese terms; the third the translation probability of the Chinese phrase to the English phrase; and the fourth vice versa.

Table 6: Bilingual Multi-word Term Examples

English Term	**Chinese Term**	**CN->EN**	**EN->CN**
AC converter	AC 转换器	0.67	0.71
AC input	AC 输入	0.33	1
AC load	AC 交流负载	1	0.14
Acceptance message	接受 消息	1	0.25
access ability	接入 能力	0.33	0.5
access API	接入 API	1	1
Access apparatus	接入 装置	0.5	0.55

We sampled 2,000 patents to evaluate the performance of these extracted pairs. In sum, 30,224 phrasal pairs of *EBil* occurs in the 2,000 patents; 24,458 (80.9% of 30,224) of these pairs also exist in *ECBil*. We ask two Chinese-English bilingual annotators to mark each of these pairs as *correct* or *wrong* according to the correctness and termhood of the translation pairs. Here we explain the term annotation standard. For the wrong cases, there are mainly three kinds: 1) wrong word/phrase pairs with irrelevant meanings; 2) verb phrases, e.g. *module locates/模块会维持*; 3) uncommon paraphrasing which likely may lead to errors if the translation is used directly, for example, *high voltage/电压高.*

For the correct cases, there are the following kinds: 1) apparently correct word/phrase pairs; 2) Chinese translations which, though missing specific details, are generally understood to refer to the English one. For example, both *high voltage/高电压* and *high voltage/高压* are considered correct; 3) common general *adjective* + *noun* phrase, where the final noun phrase is likely used in a context with specific technical meaning, e.g. *simple form/简单形式*; 4) Some terms which can be a verb or a noun depending on the context, followed by a noun phrase, e.g. *access routers/接入路由器*; 5) Some uncommon form of Chinese translations, which may sound appropriate in certain contexts but otherwise in other situations, e.g. *active application/当*

前应用; 6) Chinese translations containing Roman letters which cannot be translated, for example, *V-shaped/V 型*.

Totally, 81.2% of the 30,224 pairs in *EBil* are marked as *correct*, while 85.2% of the 24,458 pairs in *ECBil* are marked as *correct*. From these two figures, we can know that:

1) the framework used in this study can find correct bilingual terms, whose precision is higher than 80%;
2) English noun phrases are good indicators for technical terms, and it alone can achieve the precision of 81.2%;
3) Filtering with Chinese term candidates, including noun and verb phrases, can further improve the precision of extracted bilingual term candidates from 81.2% to 85.2%, but the recall drops from 100% to 80.9%.

7 Identification of Correct bilingual terms

In this section, we investigate whether we can improve the precision of extracted bilingual multi-word terms by machine learning techniques. An SVM classifier is used to help distinguish between non-terms and terms. To build the classifier, we first need to find the useful features for the differentiation of correct term and wrong terms. The features can be categorized as linguistic features and statistic features. Here we introduce the features used by our SVM classifier.

1) Linguistic features

Chinese monolingual term candidates (CMC): a binary feature indicating whether the Chinese part of the bilingual pair is a term candidate (Chinese noun/verb phrases mentioned in Sec. 3).

2) Statistic features (the first three features are got by using Moses)

Lexical weighting probability (LWP) (Koehn et al., 2007): the probability of lexical translations $\varphi(c,e)$. The formula is as follows:

$$\varphi(c,e) = \log(lex(\mathrm{e} \mid c)) + \log(lex(\mathrm{c} \mid e))$$

where $lex(\mathrm{e} \mid c)$ and $lex(\mathrm{c} \mid e)$ are the lexical weights.

CN->EN phrase translation probability (CEP) (Koehn et al., 2007): a numeric feature ranging from 0 to 1;

EN->CN phrase translation probability (ECP): similar with CEP, but the translation direction is reversed;

Frequency ratio (FR): ratio between lower and higher frequency of phrases in the pair:

For the SVM classifier, we use LIBSVM (Chang and Lin, 2001), and 5-fold cross-validation is used. Since the data is unbalanced, we train classifiers with penalty *n* for the class *wrong* and penalty 1 for class *correct*, and here we tried 1/2/2.5/3/5 for *n*. The performance comparison is shown in Table 7, from which we can observe:

1) The SVM classifier can improve the precision significantly from 0.812 to 0.908 with penalty 5 for the class wrong and penalty 1 for the class correct; while the recall drops.
2) the SVM classifier with penalty 2.5 for the class wrong outperforms ECBil at both precision and recall, showing that the features are useful for the identification of correct bilingual terms.

Table 7: SVM Classifier's Performance on Bilingual Multi-word Term Candidates

	Precision	Recall	F1
EBil	0.812	**1.000**	0.896
ECBil	0.852	0.809	0.830
1:1	0.817	0.995	**0.897**
2:1	0.847	0.911	0.878
2.5:1	0.865	0.860	0.862
3:1	0.876	0.789	0.830
5:1	**0.908**	0.617	0.735

In the following part, we investigate the contribution of each feature on the overall performance. First we build a balanced data set by using all the *wrong* pairs and the same number of *correct* pairs. The features are evaluated and the performances are shown in Table 8. We can observe that:

1) CMC is the best one among individual features, followed by LWP.
2) Combining CMC and LWP achieves good performance.
3) Combining all features outperforms individual features.

Table 8: SVM performance for feature combinations

	Precision	Recall	F1
CMC	0.581	0.882	0.701
LWP	0.673	0.665	0.669
CEP&ECP	0.550	0.825	0.660
FR	0.531	0.711	0.608
CMC&LWP	0.681	0.737	0.708
All	0.688	0.743	0.715

8 Discussion

Here we discuss some problems for bilingual term extraction. The first problem here is the definition of terminology, which has not been commonly agreed on. Thus for a word/phrase, different persons may have different answers towards the question whether it is a term. For example, our annotators disagree on whether the following three phrases are terms or not: 1)*有机无极性溶液/organic non-polar solvent*, 2) *专用扳手/ special wrench*, 3) *mask/掩模.*

The second problem here is loose translation or rewriting of Chinese phrases or paraphrasing. Many non-literal translations can be found in comparable patents. For example, *inner ring* means 内环 in Chinese, but instead *启闭环* is used in the corresponding Chinese patent. The literal translation of *非线性非树状选单* is *non-linear and non-tree-shaped menu*, but *OSD menu* is used in the corresponding English patents.

Another problem is the ambiguous words/phrases. For example, the English words/phrases in the following translation pairs are polysemous: *top block/上横梁, frame/框体, shutter/百叶门,* and they are quite ambiguous when considering independently without context. Thus it would be a little difficult to judge whether the translations are correct or not.

9 Conclusion and Future Work

In this paper, we present a framework to extract bilingual terms in comparable patents corpus. The framework includes the following major steps: 1) extract monolingual single-word and multi-word term candidates; 2) Find parallel sentences and word translations in comparable patents; 3) extract bilingual single-word and multi-word term candidates; 4) identify correct bilingual terms using a SVM classifier by integrating both linguistic and statistical information. Experiment results show that the framework can well identify bilingual terms and the SVM classifier can further improve its performance.

We would like to try more linguistic and statistic features to improve the performance of identifying correct bilingual terms. Also, monolingual term extraction and the mutual influence of monolingual and bilingual term extraction will also be examined in future. The usefulness of the extracted bilingual terms and their influence on SMT is also our interest.

References

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19(2)*:263-311.

Chang, Chih-Chung and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines.

Chen, Stanley F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. *Proceedings of ACL*, pp. 9-16. Columbus, OH.

Cheung, Lawrence, Tom Lai, Robert Luk, Oi Yee Kwong, King Kui Sin and Benjamin K. Tsou. 2002. Some Considerations on Guidelines for Bilingual Alignment and Terminology Extraction. *Proceedings of the first SIGHAN workshop on Chinese language*.

Daille, Béatrice, Éric Gaussier and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th conference on Computational linguistics*.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR)*. pp. 389-400. Tokyo, Japan.

Gaussier, Éric. 2001. General considerations on bilingual terminology extraction. *Recent Advances in Computational Terminology.*

Higuchi, Shigeto, Masatoshi Fukui, Atsushi Fujii and Tetsuya Ishikawa. 2001. PRIME: A System for Multi-lingual Patent Retrieval. In *Proceedings of MT Summit VIII*, pp. 163-167.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL, demonstration session, Prague, Czech Republic.

Lu, Bin, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang and Oi Yee Kwong. 2009. The Construction of A Chinese-English Patent Parallel Corpus. *MT Summit XII 3rd Workshop on Patent Translation*, Ottawa, Canada.

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genova, Italy.

Moore, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. *Proceedings of AMTA*, pp.135-144.

Och, Franz Joseph and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29(1)*:19-51.

Ozdowska, Sylwia. 2004. Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora. *COLING 04 Workshop on Multilingual Linguistic Resources*.

Piperidis, Stelios and Ioannis Harlas. 2006. Mining Bilingual Lexical Equivalences Out of Parallel Corpora. Lecture Notes in Computer Science, Advances in Artificial Intelligence

Toutanova, Kristina and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

Toutanova, Kristina, Dan Klein, Christopher Manning and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp.252-259.

Utiyama, Masao and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. *Proceedings of ACL*, pp. 72–79.

Utiyama, Masao and Hitoshi Isahara. 2007. A Japanese- English Patent Parallel Corpus. *MT Summit XI*, pp. 475–482.

Vintar, Špela. 2001. Using parallel corpora for translation-oriented term extraction. *Babel Journal.*

Wu, Dekai and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus, In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Wu, Dekai and Xuanyin Xia. 1995. Large-scale Automatic Extraction of an English-Chinese Translation Lexicon. *Machine Translation, 9 (3-4):* 285—313.

Zhu, Qibo, Diana Inkpen and Ash Asudeh. 2007. Automatic extraction of translations from web-based bilingual materials. *Machine Translation*.