

# Extracting Chinese-English Bilingual Core Terminology from Parallel Classified Corpora in Special Domain

Chengzhi Zhang

1. Department of Information Management, Nanjing University of Science and Technology,  
Nanjing 210093, China

2. Institute of Scientific & Technical Information of China, Beijing 100038, China  
zhangchz@itic.ac.cn

## Abstract

*The bilingual core terminology is the key resource for bilingual terminology extraction. In this paper, the keywords lists of the document in special domain are used to extract the candidate core terminology. After the keywords extraction and termhood computation, the core terminologies are extracted from the classified corpora in special domain respectively. Then, the bilingual terminology alignment method is used to extract the bilingual core terminology from the parallel classified corpora in special domain. The experiment result shows that the proposed method can be used to extract the Chinese-English core terminology quickly and efficiently.*

## 1. Introduction

Terminology is the set of technical words or expressions used in a special domain. Terminology can represent concept, and are usually used in the fields including machine translation, information retrieval, information extraction and text categorization, etc [1]. At the same time, bilingual terminology extraction plays an important role in the bilingual dictionary compilation, bilingual Ontology construction, machine translation and cross-language information retrieval etc. The bilingual core terminology is the key resource for bilingual terminology extraction. Taking bilingual core terminology pairs as bilingual terminology seed pairs and using machine learning method, we can extract more large-scale bilingual terminology pairs based on large-scale corpus. On the other hand, there are many special terminologies in the documents in special domain. The keywords lists of documents are usually candidate core terminologies. Therefore, it is significant to make full use of existing large-scale

classified resource in special domain and extract Chinese -English bilingual core terminology.

In this paper, the keywords lists of the document in special domain are used to extract the candidate core terminology. After the keywords extraction and termhood computation, the core terminology in Chinese and English are extracted from the classified corpora in special domain respectively. Then, the bilingual terminology alignment method is used to extract the bilingual core terminology from the parallel classified corpora in special domain. The experiment result shows that the average precision of the Top-200 bilingual core terminology is about 50% and the precision in some domain is 80%. The proposed method can be used to extract the core terminology in special domain quickly and efficiently.

The rest of this paper is organized as follows. The next section reviews some related work on bilingual core terminology. In section 3, a detailed description of the proposed approach is presented. Subsequently in section 4, the authors report experiments results that evaluate the proposed approach. The paper is concluded with summary and future work directions.

## 2. Related Works

The related works include bilingual terminology extraction, termhood computation, etc.

### (1) Bilingual Terminology Extraction

There are many works in the field of bilingual terminology extraction. Most of previous works are based on the parallel or comparable corpora. Sun, Jin & Du extract bilingual terminology dictionary from English and Chinese parallel corpus automatically [2]. Daille & Morin extract French-English bilingual terminology from comparable corpus [3]. Zhang, Sun, Li et al. extract special domain bilingual dictionary from non- parallel corpus, and the result shows that the

quantity and frequency of seed words have a positive effect on the result of terms extraction [4]. Erdmann Nakayama & Hara, et al. extract bilingual terminology from a multi-lingual Web-based Wikipedia [5]. Ha, Fernandez & Mitkov et al. use mutual method to extract bilingual terminology [6].

## (2) Termhood Computation

Kyo & Bin define the termhood of candidate terminology as ‘the degree of correlation between candidate terminology and concept of a special domain’ [7]. The methods of termhood computation include TF\*IDF [8][9], C-value/NC value [10], Inter-domain Entropy (IDE) [11] and Domain Component Feature Set(DCFS) [12], etc. Ji, Lu and Li et al. extract core terminology in the domain of information technique. A core term is measured by its ability to represent a single indivisible terminology in the domain as well as its ability to form other terminology in the domain [13]. Utsuro, Kida, Tonoike, et al. calculate termhood of terminology based on Web information [14][15][16].

## (3) Others Related Works

Keywords are used as candidate terminologies in this paper. It is related to the work of Utiyama, Murata & Isahara. They use keywords as training set of terminology extraction [17]. Differing from previous works, this paper combines keywords extraction and termhood calculation to extract the Chinese-English core terminology. After the keywords extraction and termhood computation, the core terminology in Chinese and English are extracted from the classified corpora in special domain respectively. Then, the bilingual terminology alignment method is used to extract the Chinese-English bilingual core terminology from the parallel classified corpora in special domain.

## 3. Extracting Bilingual Core Terminology from Parallel Classified Corpora

### 3.1 Steps of Bilingual Core Terminology

#### 3.1.1. Extracting Chinese and English Keywords.

We Use keywords extraction module to extract keywords from the real and large-scale Chinese and English classified corpora. The keywords of each document are words that can represent the topic or content of the document. These keywords are used to extract the candidate core terminology.

**3.1.2. Extracting Chinese and English Core Terminology.** According to special domain corpora,

the termhood of each candidate terminology is calculated. Under the control of termhood threshold, we can obtain core terminology list of each field. Termhood is computed based on the words distribution in each domain. Core terminology is extracted automatically based on Inter-domain Entropy.

#### 3.1.3. Extracting Chinese-English Core Terminology.

Using bilingual terminology alignment module, Chinese-English bilingual core terminology is generated automatically from the parallel classification corpora in special domain. Bilingual terminology alignment is based on the  $\chi^2$  statistics relationship between Chinese and English terminology.

## 3.2 Key Techniques of Bilingual Core Terminology

#### 3.2.1. Keyword Extraction.

The methods of keywords extraction can be divided into four categories, namely: The statistics method does not require the complex training process and is easy. Linguistics-based method mainly improves the quality of keywords by lexical analysis, syntactic analysis, semantic analysis and discourse analysis, etc. The method based on machine learning obtains statistical parameters through training the training data and extracts keywords. Hybrid method is the integration of methods above or to integrate some heuristic knowledge.

We use machine learning method to extract keywords from documents automatically. The related models include support vector machine model, multiple linear regression model, Logistic regression model and conditional random fields model (CRF), etc. A more detailed study can be found in [18].

#### 3.2.2. Termhood Computation.

Termhood computation is based on the fact that terminology is distributed in different domain corpora non-uniformly, while the general word is usually distributed uniformly. According to the method in [11], the Inter-domain Entropy of the term  $w_i$  is commuted by formula (1).

$$IDE(w_i) = \sum_j P_{ij} \log P_{ij} \quad (1)$$

Where  $IDE(w_i)$  is Inter-domain Entropy of the term  $w_i$ .  $P_{ij}$  is the occurrence probability of  $w_i$  in domain corpora  $j$ , and it is computed by formula (2).

$$P_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} \quad (2)$$

Where  $f_{ij}$  is the normalized word occurrence frequency of  $w_i$ ,  $f_{ij} = \frac{n_{ij}}{N_j}$ ,  $N_j = \sum_i n_{ij}$ .

According to the method in [11], the weight of  $w_i$  domain corpora  $j$  is calculated by formula (3).

$$W_{ij} = n_{ij} \times \log_2 \left[ \frac{N}{Nd_i} \right] \quad (3)$$

Where  $Nd_i = 2 \text{IDE}(w_i)$ , and  $N$  is the count of domain categories.

Then, the terminologies in special domain are sorted according to their weights. The Top- $k$  terminologies are selected as the candidate core terminologies.

### 3.2.3. Bilingual Core Terminology Alignment.

According to the co-occurrence in parallel corpora such as aligned title and aligned abstract, the relevancy of bilingual core terminologies is computed.

**Table 1. Occurrence frequency of bilingual terminology**

	word C in Chinese appear	word E in English doesn't appear	word E in English appears	
Word C in Chinese appear	$a$	$b$	$a+b$	
word C in Chinese doesn't appear	$c$	$d$	$c+d$	
	$a+c$	$b+d$	$N=a+b+c+d$	

Given terminology  $C$  in Chinese and terminology  $E$  in English, the relevancy of bilingual core terminology  $C$  and  $E$  can be computed by statistics methods including MI, Dice,  $\chi^2$  statistics, LogL value, etc. According to table 1, the formulas in these computation methods are as follows.

$$MI(C, E) = \log_2 (N * a / ((a+b) * (a+c))) \quad (4)$$

$$Dice(C, E) = 2 * a / ((a+b) * (a+c)) \quad (5)$$

$$\chi^2(C, E) = N * (a * d - b * c) / ((a+b) * (a+c) * (b+d) * (c+d)) \quad (6)$$

$$\begin{aligned} \text{LogL}(C, E) = & 2 * ( a * \log_2(a * N / ((a+b) * (a+c))) + \\ & b * \log_2(b * N / ((a+b) * (b+d))) + \\ & c * \log_2(c * N / ((c+d) * (a+c))) + \\ & d * \log_2(d * N / ((c+d) * (b+d))) ) \quad (7) \end{aligned}$$

Previous work shows that *LogL* method can compute the relevancy of two objects with low-frequency occurrence<sup>[19]</sup>. So, we use this method to align bilingual core terminology.

## 4. Experimental Results and Analysis

### 4.1 Training data

In this paper, we collect documents in special domain as a training set. We use academic documents as corpora for bilingual core terminology extraction and alignment. The document includes three parts, namely, title, abstract and keywords list. We collect about 460,000 Chinese documents and 130,000 English documents. The documents in the Chinese corpora and English corpora distribute in 23 categories. There is average 17, 638 documents in each category. The classified corpora are non-balanced. For example, Economics category and Culture & Science & Education & Sports category have a higher ratio. Their sum is up to 50% of the entire corpora while some category has only little documents such as Comprehensive category. Each category contains average 4, 733 parallel records. The alignment methods include title alignment which is the sentence-level alignment, abstract alignment which is the paragraph-level alignment and keywords list alignment which can be further processed to word-level alignment. The classified parallel corpora are also non-balanced.

### 4.2 Extraction Results and Analysis

The average precision of the Top-10, Top-200 and Top-500 bilingual core terminology is about 56%, 54% and 45% respectively.

The average precision of the Top-200 bilingual core terminology in the domain  $F$  is 81%. Categories like domain  $G$ ,  $B$ ,  $I$  have higher precision than categories like  $V$ ,  $X$ ,  $Z$ . According to the experiment results, we can see that the quantity of training documents has a great effect on the bilingual core terminology alignment. If the quantity is larger, the quality is higher and vice versa.

Through the result examination, we find that some core terminology which contains inclusion relations needs to be filtered. For example, the following alignment results appear in domain  $J$ , namely, Art category:

音乐教育 *music education*  
音乐 *music*

This fact suggests that it is worthy to use the method to solve the problem of inclusion relations. C-value/NC value is a good method in solving the problem. So, we plan to integrate the Inter-domain Entropy method with C-value/NC value method to extract core terminology in the future works.

## 5. Conclusions and Future Works

In this paper, the keywords lists of the document in special domain are used to extract the candidate core terminology. After the keywords extraction and termhood computation, the core terminology in Chinese and English are extracted from the classified corpora in special domain respectively. Then, the bilingual terminology alignment method is used to extract the Chinese-English bilingual core terminology from the parallel classified corpora in special domain. The experiment result shows that the proposed method can be used to extract the Chinese-English core terminology quickly and efficiently.

Our future works include: collect a relatively balanced corpus of large-scale as a training set and improve the precision of the core terminology extraction as well as bilingual core terminology alignment; integrate Inter-domain Entropy method with C-value/NC value method to extract core terminology.

## Acknowledgments

This research was partially supported by National Key Project of Scientific and Technical Supporting Programs (No.2006BAH03B02), Project of the Education Ministry's Humanities and Social Science funded by Ministry of Education of China (No.08JC870007).

## References

[1] Zhiwei Feng. A New Scientific Domain in Terminology--Computational Terminology. Terminology Standardization and Information Technology, 2008,4: 4-9.(In Chinese)

[2] Le Sun, Youbing Jin, Lin Du. Automatic Extraction of Bilingual Term Lexicon from Parallel Corpora. Journal of Chinese Information Processing, 2000, 14 (06): 33-39. (In Chinese)

[3] Béatrice Daille, Emmanuel Morin. French-English Terminology Extraction from Comparable Corpora. In: Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05), 2005: 707-718.

[4] Yongcheng Zhang, Le Sun, Fei Li, et al. Bilingual Dictionary Extraction for Special Domain Based on Web Data. Journal of Chinese Information Processing, 2006, 20 (02): 16-23. (In Chinese)

[5] Maike Erdmann, Kotaro Nakayama, Takahiro Hara and Shojiro Nishio. Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia. Journal of Information Processing, 2008, 16: 68-79.

[6] Le An Ha, Gabriela Fernandez, Ruslan Mitkov and Gloria Corpas. Mutual Bilingual Terminology Extraction. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008:

28-30.

[7] Kageura Kyo,Umino Bin. Methods of automatic term recognition: a review. Terminology, 1996, 3 (2): 259-289.

[8] Kiyotaka Uchimoto, Satoshi Sekine, Masaki Murata, Hiromi Ozaku and Hitoshi Isahara. Term Recognition by Using Different Field Corpora. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 1999: 443-450.

[9] Yirong Chen, Qin Lu, Wenjie Li, Zhifang Sui and Luning Ji. A Study on Terminology Extraction Based on Classified Corpora. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), 2006: 2383-2386.

[10] Katerina T. Frantzi, Sophia Ananiadou and Junichi Tsujii. Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, 2000, 3 (2): 115-130.

[11] Jing-Shin Chang. Domain Specific Word Extraction from Hierarchical Web Documents: A First Step toward Building Lexicon Trees from Web Corpora. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 2005: 64-71.

[12] Qinlong Zhang, Qin Lu, Zhifang Sui. Measuring Termhood in Automatic Terminology Extraction. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2007), 2007: 328-335.

[13] Luning Ji, Qin Lu, Wenjie Li, Yirong Chen. Automatic Construction of a Core Lexicon for Specific Domain. In: Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), 2007: 183-188.

[14] Mitsuhiro Kida, Masatsugu Tonoike, Takehito Utsuro, Satoshi Sato. Domain Classification of Technical Terms Using the Web. Systems and Computers in Japan, 2007, 38(14): 11-19.

[15] Takehito Utsuro, Mitsuhiro Kida, Masatsugu Tonoike, Satoshi Sato. Towards Automatic Domain Classification of Technical Terms: Estimating Domain Specificity of a Term Using the Web. In: Proceedings of Asia Information Retrieval Symposium (AIRS 2006), 2006: 633-641.

[16] Takehito Utsuro, Mitsuhiro Kida, Masatsugu Tonoike, Satoshi Sato. Collecting Novel Technical Terms from the Web by Estimating Domain Specificity of a Term. In: Proceedings of the 21st International Conference on Computer Processing of Oriental Languages (ICCPOL 2006), 2006: 173-180.

[17] Masao Utiyama, Masaki Murata and Hitoshi Isahara. Using Author Keywords for Automatic Term Recognition. Terminology, 2000, 6 (2): 313-326.

[18] Zhang Chengzhi, Wang Huilin, Yao Liu, Wu Dan, et al. Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal of Computational Information Systems, 2008, 4(3): 1169-1180.

[19] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993, 19 (1): 61-74.