The Research on the Automatic Term Extraction in the Domain of Information Science and Technology

SUI Zhifang　CHEN Yirong　HU Junfeng　WU Yunfang　YU Shiwen

(Institute of Computational Linguistics, Peking University)

szf@pku.edu.cn

Abstract: Automatic term extraction is one of the most important way to maintain and update the term bank timely. This paper puts forward a method of automatic term extraction, which extracts terms in the Domain of Information Science and Technology from large scaled domain corpus automatically. In the stage of machine learning, the statistic association information between the components of terms is obtained from corpus. The grammatical structural information inside terms is obtained from term bank. Furthermore, the domain information of the components of terms is extracted from both of term bank and corpus. In the stage of term candidates extraction, term candidates are automatically extracted using a integrated method which combine statistic-based and ruled-based methods. The preliminary experiments show, we can find a large amount of new terms using this methods, which will be of great help to the maintenance of domain term bank.

Keywords: automatic term extraction, coocurrence confidence, grammatical structural rules of term, corpus, term bank

1. Background

With the sweeping development of science and technology, new theory, new concept, new material, new technology, and new crafts turn out, meanwhile, the related terms come out and update day by day with fast speed, by extensive channels, and in large quantities, which is unprecedented. Whereas, accompanying the following problems: using of the terms confusedly for lacking coherence and standardization; one concept with multi-expressions or one expression with multi-concepts; the different meanings and misunderstanding of the neo-terms; and the different language usages between Hong Kong, Taiwan and Mainland. All of these will not only affect the understanding and communication of information which result in making the academic communication inconvenient, but also will become the barrier between China and the world. As the step of China entering the WTO and hosting the 2008 Olympic games, the criteria of the terms, the standardization of the terms and the term bank are also needed for making a bridge between China and outsides in scientific technology, economy and gym, etc.

With this purpose, the ICL worked with the national standardization institute on building of Computer-aided Term Extraction and Term Bank in the Domain of Information Technology & Science (December 2001 — December 2002). The program selects the information science and technology as an sample to formulates the related term criteria, term bank, corpus and automatic terms extraction software.

Among the above project items, automatic terms extraction is a useful way to find newly appear terms from all kinds of resources, to timely update and maintain domain term banks.

2. The state of art of automatic term extraction

The automatic term extraction related research can be divided into rule-based method【Voutilainen,1995】【Bourigault,1992】, statistic-based method【Church,1988】【Hsin-Hsi Chen, 1995】, and the hybrid method【Smadja, 1993】【Su Keh-Yih,1994】. Based on linguistic theory, the rule based one is able to treat general linguistic phenomenon. Whereas, rules are written by the human experts, so that it is hard to acquire knowledge and maintain the rule base. The statistic one obtain knowledge automatically from the corpus, so the cost of knowledge acquisition is lower. Whereas, lacking of linguistic background makes the precision lower. Hence, we use the hybrid method in the experiment of automatic term extraction.

3. The general strategy for automatic term extraction

The goal of automatic term extraction in our experiment is to extract terms in the domain of information science and technology from large corpus.

3.1 The resources used

The resources used in the automatic term extraction is as follows:

(1) The term bank in the domain of information science and technology, which includes 100 K items.
(2) The corpus in the domain of information science and technology, which includes 50,000K
(3) The software for word segmentation and part-of-speech tagging

3.2 The automatic preprocessing of the resources

3.2.1 The automatic preprocessing of the term bank

The term bank includes a large number of terms, from which we can learn the internal grammatical structural information from lexical level and syntactic level. The preprocessing of the term bank is to perform automatic word segmentation, POS tagging and extract some obvious features of the components of terms. The result after automatic word segmentation, POS tagging is as follows:

平面 plane/*n* 计算机 computer/*n* 图形学/*n* 研究 research/*v*
工具箱 tool box/*n* 实用 utility/*a-v* 程序 program/*n*
工具 tool/*n* 型 type/*k-Ng* 函数 function/*n*
工业 industry/*n* 管理 management /*v* 程序 program/*n*
工业 industry /*n* 过程 process /*n* 控制 control/*vn*
工业 industry /*n* 控制 control/*vn* 系统 system/*n*
工作 task/*n-v* 表 table/*n-Vg* 选项 option/*n* 卡 card/ *n-Ng-q-v*

In the process of POS tagging, we remain ambiguity for those words that have more than one POS. For example, “工作 task” have two POS tags, noun(*n*) and verb(*v*), it is tagged as “*n-v*”.

After word segmentation, each word in a term is called a term component. We extract some obvious features of the term component from the term bank and the corpus, which is listed as follows.

Term component	POS	First position frequency	Middle position frequency	Last position frequency

Where, “First position frequency” means the frequency of the current component in the first position of terms. “Middle position frequency” means the frequency of the current component in the middle position of terms. “Last position frequency” means the frequency of the current component in the last position of terms. In section 4.4, we will introduce how to get the above information.

3.2.2 The preprocessing of the corpus

The corpus could provide the environment information for the terms. In the preprocess stage, we perform automatic word segmentation and POS tagging. The result of the preprocess is as follows:

平面 plane/*n* 多边形 polygon /*n* 的 *de/Dg-Ng-u* 三角 triangle/*n* 剖 dissect/*v* 分 dispart /*b-n-N-Ng-q-v-Vg* 问题 problem /*n* 是 is /*r-vl* 计算 computing/*v* 几何 geometry/*n* 研究 research/*v* 的 de/*Dg-Ng-u* 一个 a/*m* 基本 basic/*a-n* 问题 problem/*n* ，/w 它 it/*r* 广泛 widely/*a* 应用 used/*v* 于 into/*p* 模式识别 pattern recognition/*tm* 、/w 图象处理 image manipulation/*tm* 、/*w* 计算机图形学 Computer Graphics/*tm* 及 and/*c-Ng-v* 机器人 automaton /*tm* 等 etc./*Ng-q-u-v* 领域 field/*n* 。/*w* 一方面 on the one hand/*c* ，/*w* 三角形 triangle/*n* 作为 as /*n-p-v* 最 most /*d-Ng* 简单 simple/*a* 的 de/*Dg-Ng-u* 平面 plane/*n* 图形 graph/*n* 较 more/*d-p* 其它 other/*r* 平面 plane/*n* 图形 graph/*n* 在 on/*d-p-v* 计算机 computer/*n* 表示 expression/*v* 、/*w* 分析 analysis/*v* 及 and/*c-Ng-v* 处理 process/*v* 时 time/*Dg-Ng-q* 方便 continent/*a-v* 得 de/*A-e-u-v* 多 a lot /*a-d-m-Ng-v* ；/*w* 另一方面 on the other hand/*c* ，/*w* 三角 triangle/*n* 剖 dissect/*v* 分 dispart/*b-n-N-Ng-q-v-Vg* 是 is/*r-vl* 研究 study/*v* 其它 other /*r* 许多 a lot of /*m* 问题 problem/*n* 的 de/*Dg-Ng-u* 前提 premise /*n* 。/*w*

Where, *tm* represents term, *a* represents adjective, *v* represents verb，*a-v* represents the current have the following two POS, “*a*”and “*v*”.

There exists a few kinds of fragments in the automatically tagged result: ordinary words such as “基本 basic ”、“ 问题 problem ” etc. Terms, such as “模式识别 pattern recognition”、”图象处理 image manipulation”、”计算机图形学 Computer Graphics”etc. The fragment of new words, such as “剖 dissect 分 dispart”. And the unknown terms, such as “三角 triangle 剖 dissect 分 dispart”, ” 计算 computing/v 几何 geometry/n”, ” 平面 plane/n 图形 graph /n” etc. Besides these, there also exists ambiguity string for word segmentation, such as “计算机系 computer department /统 system” (computer system)、”国产品 products made in China /牌 brand”(the brand of China)、”数据网 data network /络 resembling a net”(data network)、”有线电 cable eletricity /视 watch”(Cable TV). In the following, we will introduce how we repair the above errors in the automatic term extraction.

The automatic term extraction is performed on the result of the above preprocessing.

3.3 The outline of automatic term extraction

3.3.1 The definition, characteristic and representation forms of a term

- The definition of a term

According to the national standard GB／T 15237．1——2000, a term is a linguistic representation of a concept in a particular subject field.

- The characteristic of a term

According to the above definition, a term is a kind of phrase in the first place. Further more, it is different from other general phrases on that a term is usually used in a particular subject field. So, a term has two characteristics: linguistic feature and domain feature.

- The representation forms of a term

A term is a kind of phrases, whose components are close related. Further more, it has strong domain feature. The close relation of the components in a term can be captured through calculating the static association rate between the words that compose a term candidate. The linguistic feature can be captured through analysis the grammatical structural information of the terms. While the domain feature of a term can be captured through the domain component that has the possibility of

composing a term. For example, “movable terminal” and “social economy” are both composed by the close related components. While, the former is a term in the domain of information science and technology, and the latter is just a common phrase instead of a term. The reason lies in that the former has the domain feature comes from one of its components” terminal”, while the latter has not the domain feature.

In the following, we will use the above characteristic and representation forms of a term to perform automatic term extraction.

3.3.2 The outline of automatic term extraction

The outline of automatic term extraction is as follows:

- To extract close related fragments from large corpus through calculating the static association rate between the component words of the term candidates.
- To further filter the term candidates from the above results according on the grammatical structural rules of terms.
- To further filter the term candidates according on the domain feature of the components of terms.
- To construct the term bank alternatively: to select the ones that have the highest confidence to insert into the original term bank. Perform word segmentation and POS tagging to the corpus according on the new term bank, return to the first step to extract higher-level linguistic units.

4 The technology of automatic term extraction

4.1 The whole architecture of the technology of automatic term extraction

The object of the process is the automatically word segmented and POS tagged corpus. As introduced in section 3.2.2, there exist a few kinds of fragments in the automatically tagged result: ordinary words, terms, the fragment of new words, the unknown terms, and the ambiguity strings for word segmentation etc. We call each unit in the above as “word” for short. Based on such result, we evaluate the closeness of each pair of the neighboring “word”s, bind by level until the closeness of each pair of the neighboring “word”s is lower than a threshold.

The sketch map of automatic term extraction is showed in figure1:

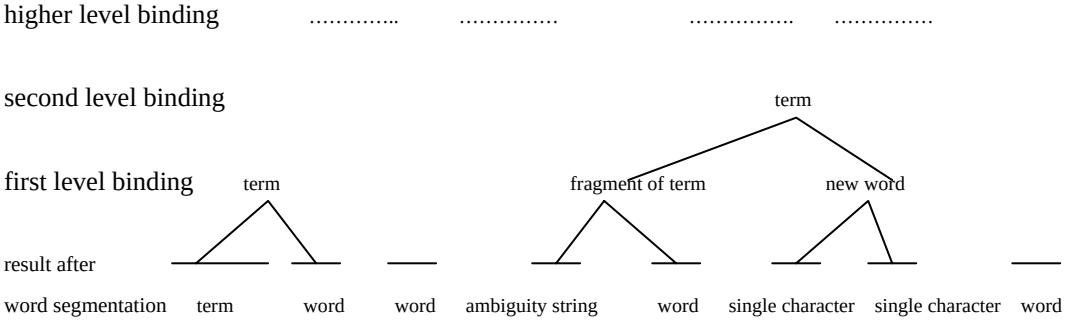

figure1: The sketch map of automatic term extraction

The flow chart of automatic term extraction is showed in figure 2.

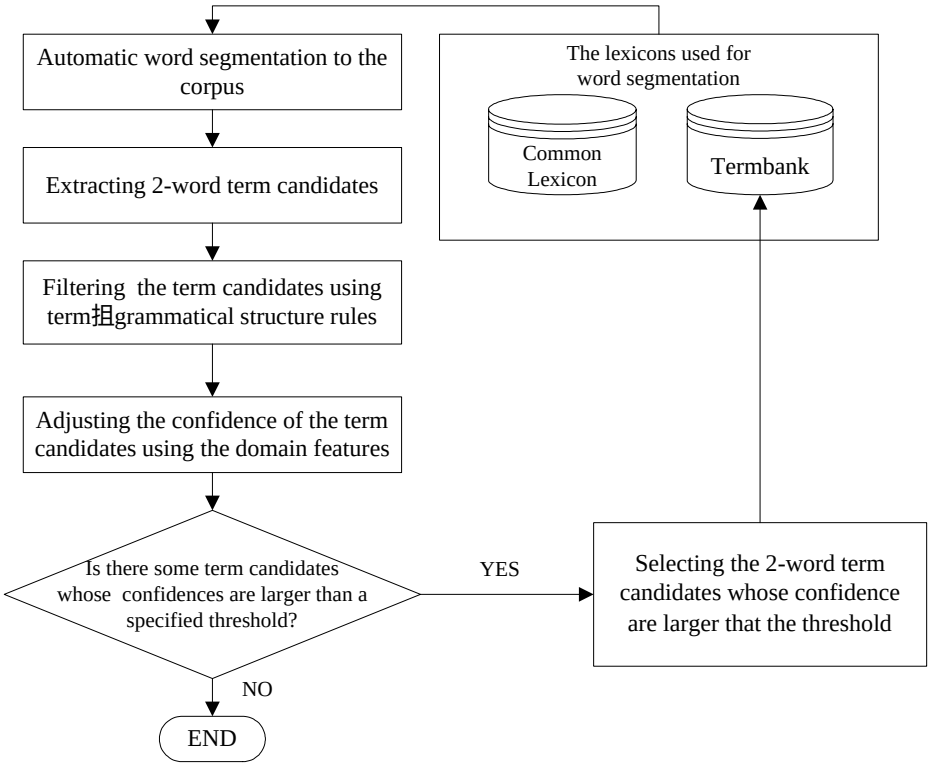

Figure 2: the flow chart of automatic term extraction is showed in

4.2 The extraction of two "word"s term candidate

Through the extraction of two "word"s term candidate, we could get the following kinds of fragments: single word terms, two words terms, two words fragments of multi-word terms, new words composed of two characters.

4.2.1 The training stage

Any of the two consecutive words W1，W2 in the corpus form a word pair（W1，W2）.The following contingency table of observed frequencies O_{ij} is set for each word pair（W1，W2）as follows：

	W2=B	W2<>B	
W1=A	O11	O12	R1
W1<>A	O21	O22	R2
	C1	C2	N

Based on the above, the following expectation table is set as follows:

	W2=B	W2<>B
W1=A	$E11=\frac{R1\times C1}{N}$	$E12=\frac{R1\times C2}{N}$
W1<>A	$E21=\frac{R2\times C1}{N}$	$E22=\frac{R2\times C2}{N}$

Where, *O11* Represents the frequency of word pair（W1，W2）in the corpus when *W1* is *A* and *W2*is *B*.
O12 Represents the frequency of word pair（W1，W2）in the corpus when *W1* is *A* and *W2* is not *B*.
O21 Represents the frequency of word pair（W1，W2）in the corpus when *W1* is not *A* and *W2* is *B*.

O22 Represents the frequency of word pair（W1，W2）in the corpus when *W1* is not *A* and *W2* is not *B*. *N* Represents the frequency of all word pair（W1，W2）in the corpus.

According to the above analysis, we could get the following equations:

O11=f(A，B)

O12=f(A)-f(A,B)

O21=f(B)-f(A,B)

O22=N-O11-O12-O21

R1=f(A)=O11+O12

R2= f(~A)=O21+O22

C1=f(B)=O11+O21

C2= f(~B)=O12+O22

N=R1+R2=C1+C2

Where, f(A,B) represent the cooccurrence of word A and word B in the corpus, f(A) represent the occurrence of word A in the corpus, f(B) represent the occurrence of word B in the corpus, f(~A) represent the occurrence of the word when the word is not A in the corpus, f(~B) represent the occurrence of the word when the word is not B in the corpus.

4.2.2 The testing stage

To any of the two words in the testing corpus, we calculate the “confidence” of the two words of forming a two-word term.

$$Confidence(w1,w2) = 2\sum_{ij} Oij \log \frac{Oij}{Eij}$$

4.2.3 The solution to data sparseness problem

The problem of data sparseness is one of the problems to deal with in the corpus based methods. Here, we use the following simple method to solve the problem.

$$O_{11}^{*} = O_{11} + 0.1$$

$$O_{12}^{*} = O_{12} + 0.1$$

$$O_{21}^{*} = O_{21} + 0.1$$

$$O_{22}^{*} = O_{22} + 0.1$$

$$E_{11}^{*} = \frac{(O_{11}^{*} + O_{12}^{*}) \times (O_{11}^{*} + O_{21}^{*})}{N}$$

$$E_{12}^{*} = \frac{(O_{11}^{*} + O_{12}^{*}) \times (O_{12}^{*} + O_{22}^{*})}{N}$$

$$E_{21}^{*} = \frac{(O_{21}^{*} + O_{22}^{*}) \times (O_{11}^{*} + O_{21}^{*})}{N}$$

$$E_{22}^{*} = \frac{(O_{21}^{*} + O_{22}^{*}) \times (O_{12}^{*} + O_{22}^{*})}{N}$$

$$Confidence^{*}(w1, w2) = 2\sum_{ij} Oij^{*} \log \frac{Oij^{*}}{Eij^{*}}$$

4.3 The filtering of term candidates by the grammatical structural rules of terms

In this stage, the term candidates extracted is filtered by the grammatical structural rules of terms. In the current stage, the grammatical structural rules are mainly obtained from the following two sources:

- ✓ The word segmented and POS tagged term set
- ✓ "Xian Dai Shu Yu Xue Yin Lun" 【Feng Zhiwei，1997】(The introduction to the contemporary terminology) written by Prof. Feng Zhiwei

In the grammatical structural rules, when a component of a term candidate has more than one POS tags, the two tags are remained in the rules. The grammatical structural rules are showed as follows:

2-word term	3-word term	4-word term	5-word term	6-word term
n+v	n+n+n	v+n+n+n	v+v+n+n+n	n+n+c+vn+n+n
(n-v)+n	n+v+n	v+n+v+n	d+v+n+n+n	n+n+vn+c+vn+n
v+n	v+v+n	n+v+v+n	m+v+m+n+n	n+n+u+b+vn+n
a+n	n+v+n	v+v+n+n	b+v+n+v+n	vn+n+vn+c+vn+n
d+n	b+v+n	(n-v)+n+(n-v)+n	n+n+v+n+n	l+vn+k+n+vn+n
b+n	n+m+n	v+n+b+n	a+n+v+n+n	n+vn+u+n+vn+n

In the above table, the POS tags in the bracket represent that the current has more than one POS tags.

After filtering according on the grammatical structural rules and the stop list, the following kinds of the term candidates extracted in section 4.2 is eliminated:

- The term candidates are common collocation, while not a complete phrase. Such as:

Commonly used collocations
为了 (for) /进一步(further) 已经(already)/被(by) 由(by)/国家(country) 具有(have)/强大(strong) 并(and)/取得(get) 中(in)/使用(use) 上(on)/运行(run) 的(de)/比重(proportion) 对(to)/这些(these) 还(yet)/将(will) 如(such as)/图(figure) 特别(especially)/是(is) 适用(fit)/于(for) 是(is)/一(a) 就(at once)/可以(can) 尤其(especially)/是(is) 分别(separately)/为(is)

4.4 The filtering of term candidates by the domain feature of the component of terms

In this stage, the term candidates are reordered according on the position information of the component of terms.

$$NewConfidence(w_1, w_2, \cdots, w_n)$$

$$= Confidence(w_1, w_2, \cdots, w_n) \times$$

$$\sqrt[n]{P(w_1 \text{ in the first position of term}) \times \ldots \times P(w_i \text{ in the mid position of term}) \times \ldots \times P(w_n \text{ in the last position of term})}$$

Where:

$$P(w \text{ in the first position of term}) =$$

$$\frac{0.8 \times w \text{ in the first position of term} + 0.1 \times w \text{ in the mid position of term} + 0.1 \times w \text{ in the last position of term}}{\text{the total number of } w \text{ in the corpus}}$$

$$P(\textit{win the mid position of term}) = \frac{0.8\times \textit{win the mid position of term} + 0.1\times \textit{win the first position of term} + 0.1\times \textit{win the last position of term}}{\textit{the total number of win the corpus}}$$

$$P(\textit{win the last position of term}) = \frac{0.8\times \textit{win the last position of term} + 0.1\times \textit{win the first position of term} + 0.1\times \textit{win the mid position of term}}{\textit{the total number of win the corpus}}$$

After filtering on the above position information of the component, the following kinds of term candidates are eliminated:

- Common phrases

The phrases useful for retrieval	The phrases not useful for retrieval
自动　柜员机 (ATM)	我们　希望 (we hope)
半导体　市场 (semiconductor market)	很　简单 (very simple)
持续　时间(persisting time)	文件　中 (in the file)
环境　保护 (environment protection)	配备　了 (is equipped)
业务　经营者 (trade operator)	转化　成 (transfer into)
基础　设施 (fundamental establishment)	是　中国 (is China)
高层　领导 (higher level leader)	刚刚　开始 (just begin)
社会　经济 (social economy)	供　选择 (to be selected)
18　微米 (18 micron)	这　项 (this item)
抢占　市场 (occupy market)	放　在 (put into)
功能　齐全 (complete function)	该　系统 (this system)

5 The analysis to the experiment result

The experiment has extracted 80 thousands term candidates from 50,000K corpus in the domain of information science and technology. For lacking of time, we have not done complete evaluation on precision and recall. The preliminary analysis to the experiment result shows that using this method, we could find a large number of new term, which is very helpful to the update of domain term bank.

6 The work in the next step

The automatic term extraction introduced in this paper is middle stage report. In the latter stage of the project, we will optimize the extraction algorithm, including: using contextual information and the template of the terms to recognize the terms.

In the future, we will study how to give the definition to the terms automatically from the large corpus.

Acknowledements

Thanks very much to Mr. Quan Ruxian, Ms. Yu Xinli, Ms. Duan Huiming, Mr. Chen Yuzhong etc.. All of them have given us great help in our research.

Reference：

1. 【Feng Zhiwei，1997】Feng Zhiwei，Xian Dai Shu Yu Xue Yin Lun" (The introduction to the contemporary terminology)，YuWen Publishing company, 1997

2. Chinese national standard GB／T 15237.1—2000
3. 【Voutilainen,1993】 Voutilainen A. Nptool, a detector of English noun phrases, In: Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Ohio State University, Columbus, USA, 1993, 48-57
4. 【Bourigault,1992】 Bourigault D. Surface grammatical analysis for extraction of terminological noun phrases, In: Proceedings of the 15th International Conference on Computational Linguistics, COLING-92, Vol-3, Nantes, France, 1992, 977-981
5. 【Church,1988】 Church K., A stochastic parts program and noun phrase parser for unrestricted text, In: Proceedings of the Second Conference on Applied Natural Language Processing, 1988
6. 【Hsin-Hsi Chen, 1995】Hsin-Hsi Chen, Development of a partially bracketed corpus with part-of-speech information only, In: Proceedings of the Fourth Workshop on Very Large Corpus, 1995
7. 【Smadja,1993】 Smadja F. Retrieving collocations from text: Xtract, In: Computational Linguistics, 19(1), 143-177

【Su Keh-Yih,1994】Su Keh-Yih, Wu Ming-Wen, A corpus-based approach to automatic compound extraction, In: Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, USA, 1994. 242-247