A Novel Parsing-based Automatic Domain Terminology Extraction Method

Ying Liu[1,2] [0000-0001-6005-5714] , Tianlin Zhang[1,2][0000-0003-0843-1916], Pei Quan[1,2], Yueran Wen[3], Kaichao Wu[4], Hongbo He[4]

[1] School of Computer and Control, University of Chinese Academy of Sciences, Beijing, 100190 China
[2] Key Lab of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, 100190 China
[3]School of Labor and Human Resources, Renmin University of China, Beijing, 100872 China
[4] Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190 China

zhangtianlin668@163.com

Abstract. As domain terminology plays a crucial role in the study of every domain, automatic domain terminology extraction method is in real demand. In this paper, we propose a novel parsing-based method which generates the domain compound terms by utilizing the dependent relations between the words. Dependency parsing is used to identify the dependent relations. In addition, a multi-factor evaluator is proposed to evaluate the significance of each candidate term which not only considers frequency but also includes the influence of other factors affecting domain terminology. Experimental results demonstrate that the proposed domain terminology extraction method outperforms the traditional POS-base method in both precision and recall.

Keywords: Domain Terminology Extraction, Dependency Parsing, Multi-factor Evaluation

1 Introduction

Domain terminology refers to the vocabulary of theoretical concepts in a specific domain. People can quickly understand the development of the subject through domain terminology, which is of great significance to scientific research. However, it is unaffordable to extract domain terminology manually from the massive text collections. Therefore, automatic domain terminology extraction is in real demand in various domains.

The process flow of the existing domain terminology extraction methods can be summarized into two steps: candidate term extraction and term evaluation [1]. Firstly, the candidate term extractor extracts terms that conform to the domain conditions.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77

Secondly, the evaluation module evaluates each candidate term and filters it when necessary based on some statistical measures.

In order to enhance the accuracy of domain terms extracted, in this paper, we propose a novel parsing-based method. The contributions of this paper can be summarized as follows:

1) Dependency parsing is proposed to be utilized to generate candidate domain terms.
2) A multi-factor evaluator is proposed, which evaluates and filters the candidate terms based on the linguistics rules, statistical methods, and domain-specific term characteristics.

We evaluated the performance of our proposed domain terminology extraction method with a frequency-based POS-based term extraction method. In the experiment, our method identified plentiful of accurate candidates. The recall rate has been improved .The ranking outperformed the counterpart in precision.

2 Related Work

Some automatic terminology recognition approaches have been proposed in recent years. The existing domain terminology extraction approaches can be classified into four categories [2]:

1) Dictionary-based method. It is simple and easy to extract domain terms by matching the words with those in a domain dictionary. However, domain terminology is constantly updated so that the domain dictionaries cannot be easily maintained [3].

2) Linguistic method. It uses the surface grammatical analysis to recognize terminology [4]. However, the linguistic rules are difficult to summarize. Linguistic method may generate lots of noise when identifying terms.

3) Statistical method. It uses the statistical properties of the terms in corpus to identify potential terminologies. Some commonly used statistical methods are word frequency statistics, TF-IDF [5], C-Value [6], etc. Statistical methods may produce some meaningless string combinations [7], common words (non-terminology) and other noises.

3 Parsing-based Domain Terminology Extraction Method

In this paper, we propose to use Dependency Parsing in the process of candidate domain term identification. The proposed parsing-based domain terminology extraction method consists of three steps: dependency parsing establishing, candidate term generation and candidate evaluation for ranking.

We will provide the details of each step in the following sections. In order to help you better understand our ideas, a Chinese corpus will be used as the example for explanation.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77

3.1 Dependency Parsing Establishment

Dependency parsing is able to reveal the syntactic structure in a given sentence by analyzing the dependencies among the components of language units. It can well explain the relationship between the adjacent words. Typical dependency parsing methods include Graph-based [8][9] and Transition-based [10][11].

The very first step in establishing dependency parsing is word segmentation. Since the CRF(Conditional Random Field)s-based word segmentation algorithm has been proved to be one of the best segmenter [12], we adopt CRFs-based parser as our baseline word segmenter. Next, a syntactic parse tree can be generated in the mean time. The dependency parsing represents the grammatical structure and the relationship between the words. Table 1 shows an example dependency parsing.

Table 1. Dependency parsing of “边际收益等于物品的价格。”(The marginal revenue equals to the price of the item.)

Dependent relation abbreviation	(word-location, word-location)
amod(adjectival modifier)	*(收益(revenue)-2, 边际(marginal)-1)*
nsubj(nominal subject)	*(等于(is equal to)-3, 收益(revenue)-2)*
root(root node)	*(ROOT-0, 等于(is equal to)-3)*
nmod:assmod(noun compound modifier)	*(价格(the price)-6, 物品(the item)-4)*
case(case)	*(物品(the item)-4, 的(of)-5)*
obj(object)	*(等于(is equal to)-3, 价格(the price)-6)*
punct(punctuation)	*(等于(is equal to)-3, 。(.)-7)*

3.2 Candidate Term Generation

In the example sentence in the previous section,收益(revenue) is a nominal subject, and 边际(marginal) serves as an adjectival modifier of 收益(revenue). By grouping words in particular roles together, we can obtain the expected "phrases". For example, 边际收益(marginal revenue) can be regarded as a candidate domain term.

Therefore, we propose to create grammatical rules to generate phrases, which can be regarded as domain terminologies. In this paper, we propose three grammatical rules, which may be widely accepted by different domains: Noun + Noun, (Adj | Noun) + Noun, and ((Adj | Noun) + (Adj | Noun)*(NounPrep)?)(Adj | Noun)*)Noun.

3.3 Candidates Evaluation and Ranking

It is inevitable that the candidate terms generated in section 3.2 may have noise. So, in order to control the quality of the selected domain terminology, we propose a set of measures in candidate evaluation. The candidates are ranked in descending order by the evaluation score for the purpose of filtering.

3.3.1 Linguistic Rule based Filter

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77

In this paper, we propose to filter the candidate terms in a “backward” manner, which filters out those candidate terms that obviously cannot be terminologies by checking with the POS of the candidate terms. Word segmentation and POS tagging are performed on the candidates.

3.3.2 Multi-factor Evaluation

Traditional terminology evaluation method is based on frequency, which sorts the candidates in descending order by their frequencies in the corpus. However, as everyone knows, although frequency is an important factor, other factors, such as adhesion, etc., also play important roles in evaluation. Therefore, we propose a multi-factor evaluator. In addition to frequency, affixes (prefixes and suffixes) that often occur in phrases are considered as a factor. The affixes of hot words in a particular domain are often the same. For example, in the domain of economics, "固定成本(constant cost) ","可变成本(variable cost) "and"总成本(total cost) " all contain the suffix "成本(cost)". Table 2 shows some affixes of the hot words and the non-terms in the candidate set in the economics corpus.

Table 2. Some affixes of the hot words and non-terms in economics

hot words Prefix	hot words Suffix	non-terms Prefix	non-terms Suffix
供给(supply)	*市场(market)*	*可以(could)*	*进行(in progress)*
福利(welfare)	*价格(price)*	*处理(deal with)*	*有关(about)*
货币(currency)	*竞争(competition)*	*进行(in progress)*	*基础(base)*
平均(average)	*成本(cost)*	*十分(very)*	*重要(important)*

Based on the observations in Table 2, affixes can either bring positive or negative impacts to domain terminology. Therefore, we propose an influence factor which indicates the impact of the affixes.

Equation 1 denotes the relationship between the frequency and the influence factor of non-terminology affixes, a is adjustment threshold.

$$\alpha = \frac{f_{word}}{a} \tag{1}$$

Equation 2 denotes the relationship between the average frequency and the influence factor of the hot-word affixes. The number of the candidate terms which occurs only once, $C_{(1)}$, is excluded, b is adjustment threshold.

$$\beta = \lceil b \frac{\sum_{i=2}^{n} f_{(i)}}{C - C_{(1)}} \rceil \tag{2}$$

Equation 3 denotes the relationship between the frequency and other factors, named as the evaluation score.

$$v = f_{word} - \alpha + \beta \tag{3}$$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77

The candidates are ranked in descending order by their evaluation scores. The higher the value, the more consistent with the characteristics of the domain terminology. By experiment, when a is 1/2 and b is 2, the effect is the best. The notations in Equation 1-3 are listed in Table 3.

Table 3. Notations used in the multi-factor evaluator

Notation	Indication
f_{word}	*The frequency of a candidate term*
α	*The influence factor of non-terminology affixes*
$f_{(i)}$	*The sum of the frequencies of the words with frequency i*
$C_{(1)}$	*The total number of candidate terms each occurring only once*
C	*The total number of the candidate terms*
β	*The influence factor of hot word affixes in a given domain*
v	The evaluation score

4 Performance Analysis

4.1 Datasets and Experiments settings

For the purpose of evaluation, we use the well-known textbook Macroeconomics (Chinese Edition) [13] as the corpus, whose domain terminology has already been labeled by domain experts. The total number of the domain terms labeled is 349.

Two different parsers are explored for comparison: Stanford parser [14], LTP parser [15]. In order to evaluate the performance of our proposed automatic parsing-based terminology extraction method, we implement the traditional POS-based method for a fair comparison. Four measures are studied in the experiments: *precision* (*P*), *recall* (*R*), *n-precision* (*P(n)*) and *n-recall* (*R(n)*) as defined in Equation 4-7. *n-precision* considers the top-n entries as well as *n-recall*.

$$P = \frac{total\ number\ of\ the\ extracted\ domain\ terms}{total\ number\ of\ extracted\ words} \times 100\% \quad (4)$$

$$R = \frac{total\ number\ of\ the\ extracted\ domain\ terms}{total\ number\ of\ the\ labeled\ doamin\ terms} \times 100\% \quad (5)$$

$$P\ (n) = \frac{total\ number\ of\ extracted\ terminologies\ in\ top{-}n\ results}{n} \times 100\% \quad (6)$$

$$R\ (n) = \frac{total\ number\ of\ extracted\ terminologies\ in\ top{-}n\ results}{total\ number\ of\ the\ labeled\ doamin\ terms} \times 100\% \quad (7)$$

4.2 Experimental Results and Discussion

Table 4 presents the precision and recall of our proposed domain terminology extraction method when using different parsers, the traditional POS, Stanford Parser and the LTP parser. The LTP parser contributes to the best precision.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77

Table 4. Total precision and recall of different methods

Method	Number of candidates	precision	recall
POS	*1117*	*12.0%*	*38.4%*
Stanford parser	*1367*	*17.6%*	*69.1%*
LTP parser	*1654*	*18.7%*	*88.5%*

In order to verify the effectiveness of the proposed multi-factor evaluator and the rationality of the ranking, n-precision and n-recall are used as the measures. The n-precision and n-recall of the extracted terms is shown in Table 5. when including the multi-factor evaluator for filtering and reordering, the n-precision rise significantly, the n-recall is higher than that of the POS-based method.

Table 5. Precision and recall of different methods in top-n results

Method	Top50 *P(n)/ R(n)*	Top100 *P(n)/ R(n)*	Top200 *P(n)/ R(n)*	Top500 *P(n)/ R(n)*
POS	*56.0%/8.0%*	*41.0%/11.7%*	*29.0%/16.6%*	*15.0%/21.5%*
Stanford parser	*50.0%/7.2%*	*25.0%/7.2%*	*19.0%/10.9%*	*11.6%/16.6%*
LTP parser	*40.0%/5.7%*	*31.0%/8.9%*	*20.5%/11.7%*	*12.4%/17.8%*
POS+ evaluator	*56.0%/8.0%*	*42.0%/12.0%*	*30.0%/17.2%*	*18.2%/26.1%*
Stanford parser +evaluator	*48.0%/6.9%*	*41.0%/11.7%*	*35.0%/20.0%*	*22.6%/32.4%*
LTP-parser + evaluator	*58.0%/8.3%*	*50.0%/14.3%*	*41.0%/23.5%*	*27.6%/39.5%*

5 Conclusion

Domain terminology is important in the study of every domain. Thus, an automatic domain terminology extraction method is in real demand. In this paper, we presented a novel automatic domain terminology extraction method. It generates the candidate domain terms by using dependency parsing. In addition, a multi-factor evaluator is proposed to evaluate the significance of each candidate term which not only considers frequency but also includes the influence of other factors affecting domain terminology. A Chinese corpus in economics is used in the performance evaluation. Experimental results demonstrate that the proposed domain terminology extraction method outperforms the traditional POS-based method in both precision and recall.

6 Acknowledgements

This project was partially supported by Grants from Natural Science Foundation of China #71671178/ #91546201/ #61202321, and the open project of the Key Lab of Big Data Mining and Knowledge Management. It was also supported by Hainan Provincial Department of Science and Technology under Grant No. ZDKJ2016021, and by Guangdong Provincial Science and Technology Project 2016B010127004.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77

Reference

1. H. Nakagawa and T. Mori, “Automatic term recognition based on statistics of compound nouns and their components,” in Terminology, pp. 201–219, 2003.
2. I. Korkontzelos, I.P. Klapaftis, and S. Manandhar, “Reviewing and evaluating automatic term recognition techniques,” in 6th international conference on Advances in Natural Language Processing, pp.248–259, 2008.
3. Krauthammer M， Nenadic G． Term Identification in the Biomedical Literature ［J］． Journal of Biomedical Informatics， 2004， 37 (6) : 512–526．
4. D. Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrases," in Proceedings of the 14th Conference on Computational Linguistics, Stroudsburg, PA, USA, 1992, pp. 977–98l.
5. Y Rezgui, "Text-based domain ontology building using tf-idf and metric clusters techniques," The Knowledge Engineering Review, vol. 22, pp. 379-403, December 2007.
6. K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," International Journal on Digital Librries, vol. 3, no. 2, pp. 115–130, 2000.
7. Damerau F J． Generating and Evaluating Domain-oriented Multi-word Terms from Texts ［J］． Information Processing ＆ Management， 1993，29(4) : 433–447．
8. Eisner, J., Three new probabilistic models for dependency parsing: An exploration, in Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 340–345, Copenhagen, URL http://cs.jhu.edu/˜jason/papers/#coling96, 1996.
9. McDonald, R., Pereira, F., Ribarov, K., and Hajic, J., Non-projective dependency parsing using spanning tree algorithms, in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 523–530, Association for Computational Linguistics, Vancouver, British Columbia, Canada, URL http://www.aclweb.org/anthology/H/H05/H05-1066, 2005.
10. Kubler, S., McDonald, R., and Nivre, J., ¨ Dependency parsing, Synthesis lectures on human language technologies, Morgan & Claypool, US, URL http://books.google.com/books?id=k3iiup7HB9UC, 2009.
11. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., and Marsi, E., MaltParser: A ¨ language-independent system for data-driven dependency parsing, Natural Language Engineering, 13, 95–135, 2007b.
12. Yang, D., Pan, Y.-c., and Furui, S. Automatic Chinese abbreviation generation using conditional random field. In NAACL '09, 273–276. 2009.
13. N. G. Mankiw, Macroeconomics, 4th ed. China Renmin University Press, 2002.
14. Dan Klein and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3–10.
15. Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13–16, Beijing, China.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_77