2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

A Chinese Domain Term Extractor

Jinbin Fu [1 +], Zhifei Wang [2] , Jintao Mao [1]

[1] Beijing Institute of Technology
[2] Harbin University

Abstract. A novel method based on statistical model of domain term language feature is proposed. Chinese domain terms have three features: domain cohesiveness, domain relevancy and domain consensus. These features are expressed respectively by statistical model and these models are integrated to extract domain terms. The relative entropy between N-Gram language models is adopt to express cohesiveness feature; the difference distributing of terms between domain corpus and balance corpus expresses the domain relevancy feature, the entropy of terms in domain corpus denotes domain consensus feature. Experimental results show this method make extraction of domain terms receiving the well precision and recall.

Keywords: term extraction, domain cohesiveness, domain relevancy, domain consensus.

1. Introduction

The vocabulary of a Chinese language contains thousands of terms, accurate identification of terms is important in a variety of contexts. Term extraction is the core parts of knowledge system and the important task in natural language processing, and it can be applied to a variety of fields such as ontology construction, text classification and information retrieval. Furthermore, we can study the development of the domain question answering system with terms.

We mainly explore the Chinese terms of computer domain in this paper, and take the 30 thousand sentences into account; the sentences come from user interactive log in a real-world web intelligent question answering system. A new method is proposed and it is based on statistical model of term language feature. Chinese domain terms have three features: domain cohesiveness, domain relevancy and domain consensus. These features are computed in respectively statistical model and these models are integrated to extract domain terms. The relative entropy between N-Gram language models is adopt to express cohesiveness feature; the difference distributing of terms between domain corpus and balance corpus expresses the domain relevancy feature, the entropy of terms in domain corpus denotes domain consensus feature. Experiments show this method make extraction of computer domain terms receiving the well precision and recall.

2. Related Work

Insofar as terms function as lexical units, their component words tend to co-occur more often, to resist substitution or paraphrase, to follow fixed syntactic patterns, and to display some degree of semantic non-compositionality [1]. However, none of these characteristics are amenable to a simple algorithmic interpretation, various term extraction systems have been developed, such as Termight [2] , and TERMS [3] among others methods [4-5] . Such systems typically rely on a combination of linguistic knowledge and statistical association measures. Grammatical patterns, such as adjective-noun or noun-noun sequences are selected then ranked statistically, and the resulting ranked list is either used directly or submitted for manual filtering. The linguistic filters are used in typical term extraction systems to reduce the number of a priori

+ Corresponding author. Tel.: +86 13910572980; fax: +86 010-68915944.
E-mail address: fujibin@gmail.com.

improbable terms and thus improve precision. The cohesiveness measure does the actual work of distinguishing between terms and plausible non-terms. A variety of methods have been applied, ranging from simple frequency [3] , modified frequency measures such as c-values [6] and standard statistical significance tests such as the t-test, the chi-squared test[7] , and log-likelihood [5] and information-based methods, e.g. point-wise mutual information [8] . These main term cohesiveness measure methods are list in Table 1.

Table 1. Term Cohesiveness Measure

Methods	formula	Interpretation
Frequency[7]	f_{xy}	f_{xy} is the frequency of the bigram xy
T-Score[7]	$\frac{f_{xy}-\frac{f_x f_y}{n}}{f_{xy}^2}$	f_x, f_y is respectively the frequency of x,y; N is sum of bigram in corpus
Log-likelihood[5]	$ll(\frac{k_1}{n_1},k_1,n_1)+ll(\frac{k_2}{n_2},k_2,n_2)$ $-ll(\frac{k_1+k_2}{n_1+n_2},k_1,n_1)-ll(\frac{k_1+k_2}{n_1+n_2},k_2,n_2)$	$k_1=f(xy),n_1=f(x^*),$ $k_2=c(\bar{xy}),n_2=f(\bar{x}^*)$ $ll(p,k,n)=k\log(p)+(n-k)\log(1-p)$
Chi-squared (x^2)[7]	$\sum_{i\in\{x,\bar{x}\}j\in\{y,\bar{y}\}}\frac{(f_{ij}-\xi_{ij})^2}{\xi_{ij}}$	$\xi_{ij}=\frac{f(i)f(j)}{N}\ i\in\{x,\bar{x}\}j\in\{y,\bar{y}\}$, $f(\bar{x})\ f(\bar{y})$ is respectively the frequency of $\bar{x}$, $\bar{y}$
Point-wise Mutual Information[8]	$\log_2\frac{p(xy)}{p(x)p(y)}$	$p(xy)$ is the frequency of the bigram xy, $p(x)$, $p(y)$ is respectively the frequency of x,y
True Mutual[5] Information	$p(xy)\log_2\frac{p(xy)}{p(x)p(y)}$	The same as above
C-Value[6]	$\begin{cases}\log_2\lvert\alpha\rvert\cdot f(a) & if\ a\ isnot\ nested\\ \text{otherwise}\\ \log_2\lvert\alpha\rvert\cdot f(a)-\frac{1}{p(T_a)}\sum_{b\in T_a}f(b)\end{cases}$	$f(\alpha)$ is the frequency of α in corpus, T_a is term candidate list including α, $P(T_a)$ is of the length of list

However, in all these studies performance was generally is very ideal, with precision falling rapidly after the very highest ranked terms list. Schone and Jurafsky [9] evaluate the identification of terms without grammatical filtering on a 6.7 million word extract from the TREC databases, applying both WordNet and online dictionaries as gold standards. Once again, the general level of performance is low, with precision falling off rapidly as larger portions of the n-best list were included, but they report better performance with statistical and information theoretic measures (including mutual information) than with frequency. The overall pattern appears to be one where lexical cohesiveness measures in general have very low precision and recall on unfiltered data, but perform far better when combined with other features which select linguistic patterns likely to function as terms.

The relatively low precision of lexical cohesiveness measures on unfiltered data no doubt has multiple explanations, but a logical candidate is the mistake of underlying statistical assumptions [7] . For instance, many of the tests assume a normal distribution, despite the highly skewed nature of natural language frequency distributions. In natural language, as first observed by Zipf [10] the frequency of words and other linguistic units tend to follow highly skewed distributions in which there are a large number of rare events. Zipf's law of this relationship for single word frequency distributions postulates that the frequency of a word is inversely proportional to its rank in the frequency distribution.

More importantly, statistical and information-based metrics such as the log-likelihood and mutual information measure significance relative to the assumption that the selection of component terms is

statistically independent. But of course the possibilities for combinations of words are not random and independent. Use of linguistic filters such as "attributive adjective + noun" or "verb + modifying prepositional phrase" arguably has the effect of selecting a subset of the language for which the standard null hypothesis -- that any word may freely be combined with any other word -- may be much more accurate, so the usual solution is to impose a linguistic filter on the data, with the cohesiveness measures being applied only to the subset thus selected. For instance, if the universe of statistical possibilities is restricted to the set of sequences in which an adjective is followed by a noun, the null hypothesis that word choice is independent -- i.e., that any adjective may precede any noun -- is a reasonable idealization. It is thus worth considering whether there are any ways to bring additional information to bear on the problem of recognizing phrasal terms without presupposing statistical independence.

3. Domain Term Extraction Based on Language Feature

In Chinese language, domain term is considered as words or phase frequently occurring in the domain corpus, expressing the concept, feature and relationship of the target domain. In term extraction area, Chinese is different from English. In Chinese, there are no obvious morphological delimiters to separate words in sentences. Hence, term extraction from Chinese is more difficult than English. We have observed that domain terms have three language features: Domain Cohesiveness, Domain Relevancy and Domain Consensus. The three features is model and integrated to evaluate domain terms.

3.1. Domain Cohesiveness

The cohesiveness measures in table 1, are mainly applied in English text, many of the measures assume a normal distribution. Furthermore, statistical and information-based metrics significance relative to the independent assumption. In the paper, we use cohesiveness measure to extract term, and filter term candidate by linguistic POS rules.

In Chinese, it is difficult to separate words by delimiters in sentences. We use cohesiveness measure to determine boundary of term. Cohesiveness of term denotes the compactness of words or characters as the component element of term. We use N-Gram language model to describe Cohesiveness degree of terms. The simplest language model is the unigram model, which assumes each word of a given word sequence is drawn independently. We denote the unigram model LM_1 for the target domain corpus. We can also train bigram models LM_2 for the corpus, it is the better model to describe two-character terms in the corpus. If we use unigram models LM_1 instead of LM_2, then we have some loss to the corpus. We assume that the amount of loss between using LM2 and LM1 is related to Cohesiveness. We use the relative entropy between Bigram model and Unigram to express the Cohesiveness of term [13]. The definition of Cohesiveness is as follow.

$$\begin{aligned} CO(W) &= \delta(LM_D^2 \mid LM_D^1) = \quad p(w)\log\frac{p(w)}{q(w)} \\ &= p(w_i \mid w_{i-1})\log\frac{p(w_i \mid w_{i-1})}{p(w_i)p(w_{i-1})} = \frac{p(w_i w_{i-1})}{p(w_{i-1})}\log\frac{p(w_i w_{i-1})}{p(w_i)p(w_{i-1})^2} \end{aligned} \tag{1}$$

Let p(x, y) is the probability of bigram of xy, occurring adjacent in the corpus. Through the cohesiveness, the adjacent characters are selected as term candidate. After above process, we can get two character term candidates. Then assuming these term candidates as a character, by re-computing of cohesiveness, these two character terms can be extend to multi-character term candidates.

The linguistic POS rules are used to filter these term candidates, it is useful to reduce the number of improbable terms and thus improve precision. The POS rules such as “adj + noun" or "noun + noun” are used as support POS rules, the POS rules such as “prepositional + verb” are used as elimination POS rules. There are some stop-words in Chinese sentence, such as “虽然”， “但是”, especially in the target domain, some general noun such as “北京” etc, is stop-words. Taking the POS rules and stop-words into account, the Domain Cohesiveness is defined as follow:

$$DCO(w) = P_{stop} \cdot P_{pos} \cdot CO(w) \tag{2}$$

While P_{stop} is a penalty factor about stop-words, P_{pos} is a penalty factor about linguistic POS rules. The term candidates are selected to next step if their domain cohesiveness value surpasses a fixed threshold, and known term list is used to determine the threshold.

3.2. Domain Relevancy

Terminological and non-terminological expression (e.g. "last week" or "real time") both have a property of high frequency in a corpus. The specificity of a terminological candidate with respect to the target domain is measured via comparative analysis across the target domain with balance corpus. Domain Relevancy of term expresses the exclusive degree of term in underlying domain.

$$DR(w) = p(w)\log\frac{p(w)}{q(w)} \tag{3}$$

Let p(w) be the probability of string w in the target corpus and q(w) be the probability of string w in the balance corpus. We can sort the term candidates by Domain Relevancy in descent order, through a threshold, relative terms in target domain can be pick out.

3.3. Domain Consensus

Terms are representative of concepts whose meaning are agreed upon large user communities in an underlying domain. We should take into account not only the overall occurrence in the target corpus but also its appearance single documents.

There are important terms with a high and average frequency within all documents in underlying domain. Distributed usage expresses a form of consensus tied to the consolidated semantics of a term within the target domain [11] . Domain Consensus measures the distributed use of a term in a domain D. The distribution of a term t in documents d_i can be taken as a stochastic variable estimated throughout all d_i in D. The entropy of this distribution expresses the consensus of t in D. The Domain Consensus is expressed as follows.

$$DC(w) = \sum_{i=1}^{m} p(w \mid d_i)\log\frac{1}{p(w \mid d_i)} \tag{4}$$

$$\text{and} \qquad p(w \mid d_i) = \frac{\text{freq}(w \text{ in } d_i)}{\sum_{d_i \in D}\text{freq}(w \text{ in } d_i)} \tag{5}$$

Let $p(w|d_i)$ be conditional probability expression of term w in document d_i, m be amount of document in the domain. Through Domain Consensus of terms, high quality term can be selected.

4. Architecture of Domain Term Extractor

The architecture of the domain term extractor is described in the figure 1, the sentences of domain corpus are segmented and processed POS tagger, then the result is as input to domain cohesiveness module, in support of POS rules, stop-words and known term list, the domain cohesiveness is computing, by the step, term candidates are selected to next step, in support of balance corpus, domain relevancy and domain consensus is computed, the terms in the domain is extracted in the system.

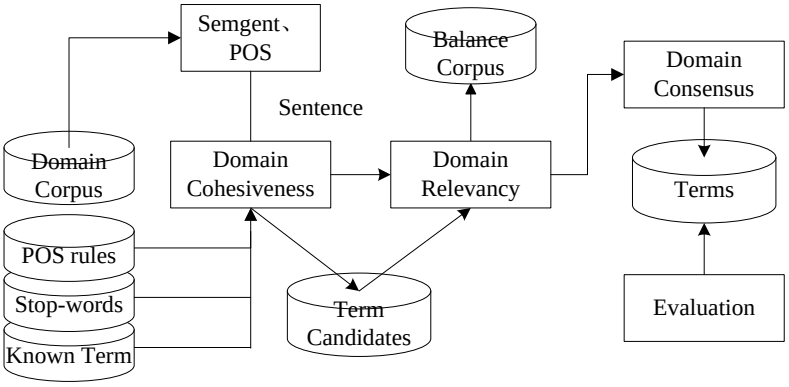

Fig. 1: Architecture of Domain Term Extractor.

5. Experiment

In the experiment, the user interactive log of a web intelligent question answering system in computer troubleshooting domain is as the domain corpus, it include 31 thousand sentence. Tancorp [12] including 14150 text files is as the balance corpus, ICTCLAS[13] is adopted as segment and POS tools , the recall and precision of the term extraction is consider as evaluation criteria. For the purpose of evaluation, the "golden standard" of domain terms is constructed manually. Typical terms extraction method such as frequency based C-Value， information theory based mutual information method is as baseline system, the result of evaluation is list in table 2. The experimental result show the method based on language feature in the paper gets better result comparing with C-value method and mutual information method.

Table 2. The evaluation of term Extraction methods

Methods	Extracted terms	Terms	Gold standard	Precision	Recall
Language Feature	254	167	235	66%	71%
C-Value	247	154	235	62%	66%
Mutual Information	235	148	235	63%	63%

6. Conclusion and Future Work

The paper presented a methods for domain extraction in computer troubleshooting domain. The method is based on statistical model of term language feature: domain cohesiveness, domain relevancy and domain consensus. Future research can focus on improving precision and recall of the extractor by linguistic knowledge.

7. References

[1] Manning C. D and H. Schutze, Foundations of Statistical Natural Language Processing. Cambridge,*MA,USA: MIT Press*, 1999.

[2] Dagan I and K. W. Church, "Termight:Identifying and translating technical terminology", *Proceedings of the fourth conference on applied natural language processing*, 1994.

[3] Justeson J. S and S. M. Katz, "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*,1995, pp. 359-371.

[4] Boguraev B and C. Kennedy, "Applications of Term Identification Technology:Domain Description and Content Characterization", *Natural Language Engineering*,1999, pp. 17-44.

[5] Patrick Pantel and Dekang Lin, "A Statistical Corpus-Based Term Extractor", Proceedings of 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: *Advances in Artificial Intelligence*, 2001.

[6] Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii , "The c-value/nc-value method of automaticrecognition for multi-word terms", *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, London, UK, 1998.

[7] Paul Deane, "A Nonparametric Method for Extraction of Candidate Phrasal Terms", *Proceedings of The 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan*, 2005.

[8] Kenneth W. Church and Patrick Hanks, "Word Association Norms, Mutual information, and Lexicography", Proceedings of the 27th. *Annual Meeting of the Association for Computational Linguistics, Vancouver, B.C.*, 1989.

[9] Patrick Schone and Daniel Jurafsky, "Is knowledge-free induction of multiword unit dictionary headwords a solved problem", *Proceedings of the Empirical Methods in Natural Language Processing*, 2001.

[10] George Kingsley Zipf, *Human Behavior and the Principle of Least Effort: Addison-Wesley*, 1949.

[11] LIU Tao, LIU Bing-quan, XU Zhi-ming, and WANG Xiao-long, "Automatic Domain-Specific Term Extraction and Its Application in Text Classification", *ACTA ELECTRONICA SINICA*,2007, pp. 328-332.

[12] Songbo Tan, A Novel Refinement Approach for Text Categorization.: *ACM CIKM*, 2005.

[13] Huaping Zhang, "Chinese Lexical Analysis Using Hierarchical Hidden Markov Model", *Proceedings of Second SIGHAN workshop affiliated with 41th ACL*, Sapporo Japan, 2003.