Preliminary Chinese Term Classification for Ontology Construction

Gaoying Cui, Qin Lu, Wenjie Li
Department of Computing,
Hong Kong Polytechnic University
{csgycui, csluqin, cswjli}@comp.polyu.edu.hk

Abstract

An ontology can be seen as a representation of concepts in a specific domain. Accordingly, ontology construction can be regarded as the process of organizing these concepts. If the terms which are used to label the concepts are classified before building an ontology, the work of ontology construction can proceed much more easily. Part-of-speech (PoS) tags usually carry some linguistic information of terms, so PoS tagging can be seen as a kind of preliminary classification to help constructing concept nodes in ontology because features or attributes related to concepts of different PoS types may be different. This paper presents a simple approach to tag domain terms for the convenience of ontology construction, referred to as *Term PoS* (*TPoS*) *Tagging*. The proposed approach makes use of segmentation and tagging results from a general PoS tagging software to predict tags for extracted domain specific terms. This approach needs no training and no context information. The experimental results show that the proposed approach achieves a precision of 95.41% for extracted terms and can be easily applied to different domains. Comparing with some existing approaches, our approach shows that for some specific tasks, simple method can obtain very good performance and is thus a better choice.

Keywords: ontology construction, part-of-speech (PoS) tagging, Term PoS (TPoS) tagging.

1 Introduction

Ontology construction has two main issues including the acquisition of domain concepts and the acquisition of appropriate taxonomies of these concepts. These concepts are labeled by the terms used in the domain which are described by different attributes. Since domain specific terms (terminology) are labels of concepts among other things, terminology extraction is the first and the foremost important step of domain concept acquisition. Most of the existing algorithms in Chinese terminology extraction only produce a list of terms without much linguistic information or classification information (Yun Li and Qiangjun Wang, 2001; Yan He et al., 2006; Feng Zhang et al., 2006). This fact makes it difficult in ontology construction as the fundamental features of these terms are missing. The acquisition of taxonomies is in fact the process of organizing domain specific concepts. These concepts in an ontology should be defined using a subclass hierarchy by assigning and defining properties and by defining relationship between concepts etc. (Van Rees, R., 2003). These methods are all concept descriptions. The linguistic information associated with domain terms such as PoS tags and semantic classification information of terms can also make up for the concept related features which are associated with concept labels. Terms with different PoS tags usually carry different semantic information. For example, a noun is usually a word naming a thing or an object. A verb is usually a word denoting an action, occurrence or state of existence, which are all associated with a time and a place. Thus in ontology construction, noun nodes and verb nodes should be described using different attributes with different discriminating characters. With this information, extracted terms can then be classified

accordingly to help in ontology construction and retrieval work. Thus PoS tags can help identify the different features needed for concept representation in domain ontology construction.

It should be pointed out that *Term PoS* (*TPoS*) tagging is different from the general PoS tagging tasks. It is designed to do PoS tagging for a given list of terms extracted from some terminology extraction algorithms such as those presented in (Luning Ji et al., 2007). The granularity of general PoS tagging is smaller than what is targeted in this paper because terms representing domain specific concepts are more likely to be compound words and sometimes even phrases, such as "文件管理器"(file manager), "并发描述"(description of concurrency), etc.. Even though current general word segmentation and PoS tagging can achieve precision of 99.6% and 97.58%, respectively (Huaping Zhang et al., 2003), its performance for domain specific corpus is much less satisfactory (Luning Ji et al., 2007), which is why terminology extraction algorithms need to be developed.

In this paper, a very simple but effective method is proposed for TPoS tagging which needs no training process or even context information. This method is based on the assumption that every term has a headword. For a given list of domain specific terms which are segmented and each word in the term already has a PoS tag, the TPoS tagging algorithm then identifies the position of the headword and take the tag of the headword as the tag of the term. Experiments show that this method is quite effective in giving good precision and minimal computing time.

The remaining of this paper is organized as follows. Section 2 reviews the related work. Section 3 gives the observations to the task and corresponding corpus, then presents our method for TPOS tagging. Section 4 gives the evaluation details and discussions on the proposed method and reference methods. Section 5 concludes this paper.

2 Related Work

Although TPoS tagging is different from general PoS tagging, the general POS tagging methods are worthy of referencing. There are a lot of existing POS tagging researches which can be classified into following categories in general. Natural ideas of solving this problem were to make use of the information from the words themselves. A number of features based on prefixes and suffixes and spelling cues like capitalization were adopted in these researches (Mikheev, A, 1997; Brants, Thorsten, 2000; Mikheev, A, 1996). Mikheev presented a technique for automatically acquiring rules to guess possible POS tags for unknown words using their starting and ending segments (Mikheev, A, 1997). Through an unsupervised process of rule acquisition, three complementary sets of word-guessing rules would be induced from a general purpose lexicon and a raw corpus: prefix morphological rules, suffix morphological rules and ending-guessing rules (Mikheev, A, 1996). Brants used the linear interpolation of fixed length suffix model for word handling in his POS tagger, named TnT. For example, an English word ending in the suffix *–able* was very likely to be an adjective (Brants, Thorsten, 2000).

Some existing methods are based on the analysis of word morphology. They exploited more features besides morphology or took morphology as supplementary means (Toutanova et al., 2003; Huihsin Tseng et al., 2005; Samuelsson, Christer, 1993). Toutanova et al. demonstrated the use of both preceding and following tag contexts via a dependency network representation and made use of some additional features such as lexical features including jointly conditioning on multiple consecutive words and other fine-grained modeling of word features (Toutanova et al., 2003). Huihisin et al. proposed a variety of morphological word features, such as the tag sequence features from both left and right side of the current word for POS tagging and implemented them in a Maximum Entropy Markov model (Huihsin Tseng et al., 2005). Samuelsson used n-grams of letter sequences ending and starting each word as word features. The main goal of using Bayesian inference was to investigate the influence of various information sources, and ways of combining them, on the ability to assign lexical categories to words. The Bayesian inference was used to find the tag assignment T with highest probability $P(T|M, S)$ given morphology M (word form) and syntactic context S (neighboring tags) (Samuelsson, Christer, 1993).

Other researchers inclined to regard this POS tagging work as a multi-class classification problem. Many methods used in machine learning, such

as Decision Tree, Support Vector Machines (SVM) and *k*-Nearest-Neighbors (*k*-NN), were used for guessing possible POS tags of words (G. Orphanos and D. Christodoulakis, 1999; Nakagawa T, 2001; Maosong Sun et al., 2000). Orphanos and Christodoulakis presented a POS tagger for Modern Greek and focused on a data-driven approach for the induction of decision trees used as disambiguation or guessing devices (G. Orphanos and D. Christodoulakis, 1999). The system was based on a high-coverage lexicon and a tagged corpus capable of showing off the behavior of all POS ambiguity schemes and characteristics of words. Support Vector Machine is a widely used (or effective) classification approach for solving two-class pattern recognition problems. Selecting appropriate features and training effective classifiers are the main points of SVM method. Nakagawa et al. used substrings and surrounding context as features and achieve high accuracy in POS tag prediction (Nakagawa T, 2001). Furthermore, Sun et al presented a POS identification algorithm based on k-nearest-neighbors (*k*-NN) strategy for Chinese word POS tagging. With the auxiliary information such as existing tagged lexicon, the algorithm can find out *k* nearest words which were mostly similar with the word need tagging (Maosong Sun et al., 2000).

3 Algorithm Design

As pointed out earlier, TPoS tagging is different from the general PoS tagging tasks. In this paper, it is assumed that a terminology extraction algorithm has already obtained the PoS tags of individual words. For example, in the segmented and tagged sentence "计算机/n 图形/n 学/v 中/f ，/w 物体/n 常常/d 用/v 多边形/a 网格/n 来/f 表示/v 。/w"(In computer graphics, objects are usually represented as polygonal meshes.), the term "多边形网格" (polygonal meshes) has been segmented into two individual words and tagged as "多边形/a" (polygonal /a) and "网格/n" (meshes /n). The terminology extraction algorithm would identify these two words "多边形/a" and "网格/n" as a single term in a specific domain. The proposed algorithm is to determine the PoS of this single term "多边形网格" (polygonal meshes), thus the algorithm is referred to as TPoS tagging. It can be seen that the general purpose PoS tagging and term PoS tagging assign tags at different granularity. In principle, the context information of terms can help TPoS tagging and the individual PoS tags may be good choices as classification features.

The proposed TPoS tagging algorithm consists of two modules. The first module is a terminology extraction preprocessing module. The second module carries out the TPoS tag assignment. In the terminology extraction module, if the result of terminology extraction algorithm is a list of terms without PoS tags, a general purpose segmenter called ICTCLAS[1] will be used to give PoS tags to all individual words. ICTCLAS is developed by Chinese Academy of Science, the precision of which is 97.58% on tagging general words (Huaping Zhang et al., 2003). Then the output of this module is a list of terms, referred to as TermList, using algorithms such as the method described in (Luning Ji et al., 2007).

In this paper, two simple schemes for the term PoS tag assignment module are proposed. The first scheme is called the *blind assignment scheme*. It simply assigns the noun tag to every term in the TermList. This is based on the assumption that most of the terms in a specific domain represent certain concepts that are most likely to be nouns. Result from this blind assignment scheme can be considered as the baseline or the worse case scenario. Even in general domain, it is observed that nouns are in the majority of Chinese words with more than 50% among all different PoS tags (Hui Wang, 2006).

The second scheme is called *head-word-driven assignment scheme*. Theoretically, it will take the tag of the head word of one term as the tag of the whole term. But here it simply takes the tag of the last word in a term. This is based on the assumption that each term has a headword which in most cases is the last word in a term (Hui Wang, 2006). One additional experiment has been done to verify this assumption. A manually annotated Chinese shallow Treebank in general domain is used for the statistic work (Ruifeng Xu et al., 2005). There are 9 different structures of Chinese phrases, (Yunfang Wu et al., 2003), but only 3 of them do not have their head words in the tail, which are about 6.56% from all phrases. Following the examples earlier,

[1] Copyright © Institute of Computing Technology, Chinese Academy of Sciences

the term “多边形/a 网格/n” (polygonal /a meshes /n) will be assigned the tag “/n” because the last word is labeled “/n”.

There are a lot of semanteme tags at the end of a term. For example, “/ng” presents single character postfix of a noun. But it would be improper if a term is tagged as “/ng”. For example, the term “决策器” (decision-making machine) contains two segments as listed with two components “决策/n” and “器/ng”. It is obvious that “决策器/ng” is inappropriate. Thus the head-word-driven assignment scheme also includes some rules to correct this kind of problems. As will be discussed in the experiment, the current result of TPoS tagging is based on 2 simple induction rules applied in this algorithm.

4 Experiments and Discussions

The domain corpus used in this work contains 16 papers selected from different Chinese IT journals between 1998 and 2000 with over 1,500,000 numbers of characters. They cover topics in IT, such as electronics, software engineering, telecom, and wireless communication. The same corpus is used by the terminology extraction algorithm developed in (Luning Ji et al., 2007). In the domain of IT, two TermLists are used for the experiment. TermList1 is a manually collected and verified term list from the selected corpus containing a total of 3,343 terms. TermList1 is also referred to as the standard answer set to the corpus for evaluation purposes. TermList2 is produced by running the terminology extraction algorithm in (Van Rees, R, 2003). TermList2 contains 2,660 items out of which 929 of them are verified as terminology and 1,731 items are not considered terminology according to the standard answer above.

To verify the validity of the proposed method to different domains, a term list containing 366 legal terms obtained from Google searching results for “法律术语大全”(complete dictionary of legal terms) is selected for comparison, which is named TermList3.

4.1 Experiment on the Blind Assignment Scheme

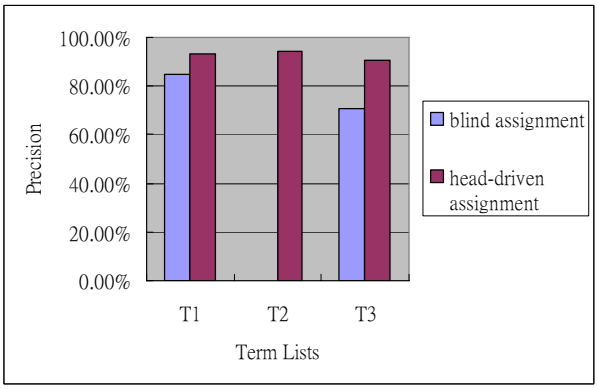
The first experiment is designed to examine the proportion of nouns in TermList1 and TermList3, to validate of the assumption of the blind assignment scheme. In first part of this experiment, all the 3,343 terms in TermList1 are tagged as nouns. The result shows that the precision of the blind assignment scheme is between 78.79% and 84.77%. The reason for the range is that there are about 200 terms in TermList1 which can be considered either as nouns, gerunds, or even verbs without reference to context. For example, the term “局域网远程访问” (“remote access of local area network” or “remote access to local area network”) and the term “极化” (polarization *or* polarize), can be considered either as nouns if they are regarded as courses of events or as verbs if they refer to the actions for completing certain work. The specific type is dependent on the context which is not provided without the use of a corpus. However, the experiment result does show that in a specific domain, there is a much higher percentage of terms that are nouns than other tags in general (Hui Wang, 2006). As to TermList3, the precision of blind assignment is between 65.57% and 70.77% (19 mixed ones). TermList2 is the result of a terminology extraction algorithm and there are non-term items in the extraction result, so the blind assignment scheme is not applied on TermList2. The blue colored bars (lighter color) in **Figure 1** shows the result of TermList1 and TermList3 using the blind assignment scheme which gives the two worst result compared to our proposed approach to be discussed in Section 4.2

4.2 Experiments on the Head-Word-Driven Assignment Scheme

The experiment in this section was designed to validate the proposed head-word-driven assignment scheme. The same experiment is conducted on the three term lists respectively, as shown in **Figure 1** in purple color (darker color). The precision for assigning TPoS tags to TermList1 is 93.45%. By taking the result from a terminology extraction algorithm without regards to its potential error propagations, the precision of the head-word-driven assignment scheme for TermList2 is 94.32%. For TermList3, the precision of PoS tag assignment is 90.71%. By comparing to the blind assignment scheme, this algorithm has reasonably good performance for all three term list with precision of over 90%. It also gives 8.7% and 19.9% improvement for TermList1 and TermList3, respectively, as compared to the blind assignment

scheme, a reasonably good improvement without a heavy cost. However, there are some abnormalities in these results. Supposedly, TermList1 is a hand verified term list in the IT domain and thus its result should have less noise and thus should perform better than TermList2, which is not the case as shown in **Figure 1**.

Figure 1 Performance of the Two Assignment Schemes on the Three Term Lists



By further analyzing the error result, for example for TermList1, among these 3,343 terms, about 219 were given improper tags, such as the term "图形学" (Graphics). In this example, two individual words, "图形/n" and "学/v", form a term. So the output was "图形学/v" for taking the tag of the last segment. It was a wrong tag because the whole term was a noun. In fact, the error is caused by the general word PoS tagging algorithm because without context, the most likely tagging of "学", a semanteme, is a verb. This kind of errors in semanteme tagging appeared in the results of all three term lists with 169 from TermList1, 29 from TermList2 and 12 from TermList3, respectively. This was a kind of errors which can be corrected by applying some simple induction rules. For example, for all semantemes with multiple tags (including noun as in the example), the rule can be "tagging terms with noun suffixes as nouns". For example, terms "劳改/n 场/q" (reform-through-labor camp) and "计算机/n 图形/n 学/v" (computer graphics) were given different tags using the head-word-driven assignment scheme. They were assigned as: "劳改场/q" and "计算机图形学/v" which can be corrected by this rule. Another kind of mistake is related to the suffix tags such as "/ng" (noun suffix) and "/vg"(verb suffix). For examples, "知识/n 产权/n 庭/ng" (intellectual property tribunal) and "数据/n 集/vg" (data set) will be tagged as "知识产权庭/ng" and "数据集/vg", respectively, which are obviously wrong. So, the simple rule of "tagging terms with "/ng" and "/vg" to "/n" is applied. The performance of TPoS tag assignment after applying these two fine tuning induction rules are shown in **Table 1** below.

Table 1 Influence of Induction Rules on Different Term Lists

Term Lists	Precision of tagging	Precision after adding induction rule	Improvement Percentage
TermList1	93.45%	97.03%	3.83%
TermList2	94.32%	95.41%	1.16%
TermList3	90.71%	93.99%	3.62%

It is obvious that with the use of fine tuning using induction rules, the results are much better. In fact the result for TermList1 reached 97.03% which is quite close to PoS tagging of general domain data. The abnormality also disappeared as the performance of TermList1 has the best result. The improvement to TermList2 (1.16%) is not as obvious as that for TermList1 and TermList3, which are 3.83% and 3.62%, respectively. This, however, is reasonable as TermList2 is produced directly from a terminology extraction algorithm using a corpus, thus, the results are noisier.

Further analysis is then conducted on the result of TermList2 to examine the influence of non-term items to this term list. The non-term items are items that are general words or items cannot be considered as terminology according to the standard answer sheet. For example, neither of the terms "问题" (problem) and "模式训练是" (pattern training is) were considered as terms because the former was a general word, and the latter should be considered as a fragment rather than a word. In fact, in 2,660 items extracted by the algorithm as terminology, only 929 of them are indeed terminology (34.92%), and rest of them do not qualify as domain specific terms. The result of this analysis is listed in **Table 2**.

Table 2 Data Distribution Analysis on TermList2

	Without Induction Rules		Induction Rules Applied	
	correct terms	precision	correct terms	precision
Terms (929)	879	94.62%	898	96.66%
Non-terms (1,731)	1,630	94.17%	1,640	94.74%
Total (2,660)	2,509	94.32%	2,538	95.41%

Results show that 31 and 50 from the 929 correct terms were assigned improper PoS tags using the proposed algorithm with and without the inductions rules, respectively. That is, the precisions for correct data are comparable to that of TermList1 (93.45% and 97.03%, respectively). For the non-terms, 91 items and 101 items from 1,731 items were assigned improper tags with and without the induction rules, respectively. Even though the precisions for terms and non-terms without using the induction rules are quite the same (94.62% vs. 94.17%), the improvement for the non-terms using the induction rules are much less impressive than that for the terms. This is the reason for the relatively less impressed performance of induction rules for TermList2. It is interesting to know that, even though the performance of the terminology extraction algorithm is quite poor with precision of only around 35% (929 out of 2,666 terms), it does not affect too much on the performance of the TPoS proposed in this paper. This is mainly because the items extracted are still legitimate words, compounds, or phrases which are not necessarily domain specific.

The proposed algorithm in this paper use minimum resources. They need no training process and even no context information. But the performance of the proposed algorithm is still quite good and can be directly used as a preparation work for domain ontology construction because of its presion of over 95%. Other PoS tagging algorithms reach good performance in processing general words. For example, a k-nearest-neighbors strategy to identify possible PoS tags for Chinese words can reach 90.25% for general word PoS tagging (Maosong Sun et al., 2000). Another method based on SVM method on English corpus can reach 96.9% in PoS tagging known and unknown words (Nakagawa T, 2001). These results show that proposed method in this paper is comparable to these general PoS tagging algorithms in magnitude. Of course, one main reason of this fact is the difference in its objectives. The proposed method is for the PoS tagging of domain specific terms which have much less ambiguity than tagging of general text. Domain specific terms are more likely to be nouns and there are some rules in the word-formation patterns while general PoS tagging algorithms usually need training process in which large manually labeled corpora would be involved. Experiment results also show that this simple method can be applied to data in different domains.

5 Conclusion and Future Work

In this paper, a simple but effective method for assigning PoS tags to domain specific terms was presented. This is a preliminary classification work on terms. It needs no training process and not even context information. Yet it obtains a relatively good result. The method itself is not domain dependent, thus it is applicable to different domains. Results show that in certain applications, a simple method may be more effective under similar circumstances. The algorithm can still be investigated over the use of more induction rules. Some context information, statistics of word/tag usage can also be explored.

Acknowledgments

This project is partially supported by CERG grants (PolyU 5190/04E and PolyU 5225/05E) and B-Q941 (Acquisition of New Domain Specific Concepts and Ontology Update).

References

Yun Li, Qiangjun Wang. 2001. *Automatic Term Extraction in the Field of Information Technology*. In the proceedings of The Conference of 20th Anniversary for Chinese Information Processing Society of China.

Yan He, Zhifang Sui, Huiming Duan, and Shiwen Yu. 2006. *Term Mining Combining Term Component Bank. In Computer Engineering and Applications*. Vol.42 No.33,4--7.

Feng Zhang, Xiaozhong Fan, and Yun Xu. 2006. *Chinese Term Extraction Based on PAT Tree*. Journal of Beijing Institute of Technology. Vol. 15, No. 2.

Van Rees, R. 2003. *Clarity in the Usage of the Terms Ontology, Taxonomy and Classification*. CIB73.

Mikheev, A. 1997. *Automatic Rule Induction. for Unknown Word Guessing*. In Computational Lingusitics Vol. 23(3), ACL.

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.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005.*Morphological Features Help POS Tagging of Unknown Words across Language Varieties*. In proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Samuelsson, Christer. 1993. *Morphological Tagging Based Entirely on Bayesian Inference.* In proceedings of NCCL 9.

Brants, Thorsten. 2000. *TnT: A Statistical Part-of-Speech Tagger*. In proceedings of ANLP 6.

G. Orphanos, and D. Christodoulakis. 1999. *POS Disambiguation and Unknown Word Guessing with Decision Trees*. In proceedings of EACL'99, 134--141.

H Schmid. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In proceedings of International Conference on New Methods in Language Processing.

Maosong Sun, Dayang Shen, and Changning Huang. 1997. *CSeg& Tag1.0: a practical word segmenter and POS tagger for Chinese texts*. In proceedings of the fifth conference on applied natural language processing.

Ying Liu. 2002. *Analysing Chinese with Rule-based Method Combined with Statistic-based Method*. In Computer Engineering and Applications, Vol.7.

Mikheev, A. 1996. *Unsupervised Learning of Word-Category Guessing Rules*. In proceedings of ACL-96.

Nakagawa T, Kudoh T, and Matsumoto Y. 2001. *Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines*. In proceedings of NLPPRS 6, 325--331.

Maosong Sun, Zhengping Zuo, and B K, TSOU. 2000. *Part-of-Speech Identification for Unknown Chinese Words Based on K-Nearest-Neighbors Strategy*. In Chinese Journal of Computers. Vol.23 No.2: 166--170.

Luning Ji, Mantai Sum, Qin Lu, Wenjie Li, Yirong Chen. 2007. *Chinese Terminology Extraction using Window-based Contextual Information.* In proceedings of CICLING.

Huaping Zhang et al. 2003. *HHMM-based Chinese Lexical Analyzer ICTCLAS*. Second SIGHAN workshop affiliated with 41th ACL, 184--187. Sapporo Japan.

Hui Wang. Last checked: 2007-08-04. *Statistical studies on Chinese vocabulary (汉语词汇统计研究)*. http://www.huayuqiao.org/articles/wanghui/wanghui06.doc. The date of publication is unknown from the online source.

Ruifeng Xu, Qin Lu, Yin Li and Wanyin Li. 2005. *The Design and Construction of the PolyU Shallow Treebank*. International Journal of Computational Linguistics and Chinese Language Processing, V.10 N.3.

Yunfang Wu, Baobao Chang and Weidong Zhan. 2003. *Building Chinese-English Bilingual Phrase Database*. Page 41-45, Vol. 4.