SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus

Shih-Hung WU
Institute of Information Science
Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
shwu@iis.sinica.edu.tw

Wen-Lian HSU
Institute of Information Science
Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
hsu@iis.sinica.edu.tw

Abstract

In this paper, we focus on the domain ontology acquisition from Chinese corpus by extracting rules designed for Chinese phrases. These rules are noun sequences with part-of-speech tags. Experiments show that this process can construct domain ontology prototypes efficiently and effectively.

1. Introduction

Domain ontology is important for large-scale natural language application systems such as speech recognition (Flett & Brown 2001), question answering (QA), knowledge management and organization memory (KM/OM), information retrieval, machine translation (Guarino 1998), and grammar checking systems (Bredenkamp 2000). With the help of domain ontology, software systems can perform better in understanding natural language. However, building domain ontology is laborious and time consuming.

Previous works suggest that ontology acquisition is an iterative process which includes keyword collection as well as structure reorganization. The ontology will be revised, refined, and filled in detail during iteration. (Noy and McGuinness 2001) For example (Hearst 1992), in order to find a hyponym of a keyword, the human editor must observe sentences containing this keyword and its related hyponyms. The editor then deduces rules for finding more hyponyms of this keyword. As such cycle iterates, the editor refines the rules to obtain better quality pairs of keyword-hyponyms. In this work we try to speed up the above labor-intensive approach by designing acquisition rules that can be applied recursively. A human editor only has to verify the results of the acquisition.

The extraction rules we specified are templates of part-of-speech (POS) tagged phrase structure. Parsing a phrase by POS tags (Abney 1991) is a well-known shallow parsing technique, which provides the natural language processing function for different natural language applications including ontology acquisition (Maedche and Staab 2000).

In previous works (Hsu et al. 2001), we have constructed a knowledge representation framework, InfoMap, to integrate various linguistic knowledge, commonsense knowledge and domain knowledge. InfoMap is designed to perform natural language understanding. It has been applied to many application domains, such as QA system and KM/OM (Wu et al. 2002) and has obtained encouraging results. An important characteristic of InfoMap is to extract events from a sentence by capturing the topic words, usually noun-verb (NV) pairs or noun-noun (NN) pairs, which is defined in domain ontology. We design the SOAT as a semi-automatic domain ontology acquisition tool following the ontology framework, InfoMap.

We shall review the InfoMap ontology framework in section 2. The domain ontology acquisition process and extraction rules will be discussed in Section 3. Experimental results are reported in section 4. We conclude our work in Section 5.

2. The InfoMap Framework

Gruber defines an ontology to be a description of concepts and relationships (Gruber 1993). Our knowledge representation scheme, InfoMap, can serve as an ontology framework. InfoMap provides the knowledge necessary for understanding natural language related to a

certain knowledge domain. Thus, we need to integrate various linguistic knowledge, commonsense knowledge and domain knowledge in making inferences.

2.1 The Structure of InfoMap

InfoMap consists of domain concepts and their associated attributes, activities, etc., which are its related concepts. Each of the concepts forms a tree-like taxonomy. InfoMap defines "reference" nodes to connect nodes on different branches, thereby integrating these concepts into a semantic network.

InfoMap not only classifies concepts, but also classifies the relationships among concepts. There are two types of nodes in InfoMap: concept nodes and function nodes. The root node of a domain is the name of the domain. Following the root node, topics are found in this domain that may be of interest to users. These topics have sub-categories that list related sub-topics in a recursive fashion.

2.2 Function Nodes in InfoMap

InfoMap uses function nodes to label different relationships among related concept nodes. The basic function nodes are: category, attribute, synonym, and activity, which are described below.

1. Category: Various ways of dividing up a concept *A*. For example, for the concept of "people", we can divide it into young, mid-age and old people according to "age". Another way is to divide it into men and women according to "sex", or rich and poor people according to "wealth", etc. For each such partition, we shall attach a "cause". Each such division can be regarded as an angle of viewing concept *A*.
2. Attribute: Properties of concept *A*. For example, the attributes of a human being can be the organs, the height, the weight, hobbies, etc.
3. Associated activity: Actions that can be associated with concept *A*. For example, if *A* is a "car", then it can be driven, parked, raced, washed, repaired, etc.
4. Synonym: Expressions that are synonymous to concept A in the context.

2.3 The Contextual View of InfoMap

Generally speaking, an ontology consists of definitions of concepts, relations and axioms. A well known ontology, WordNet (Miller 1990), has the following features: hypernymy, hyponymy, antonymy, semantic relationship, and synset. Comparing with the globlal view of concepts in WordNet, InfoMap defines category, event, atttibute, and synonym in a more contextual fashion. For example, the synonym of a concept in InfoMap is valid only in this particular context. This is very different from the synset in WordNet. Each node *B* underneath a function node (synonym, attribute, activity or category) of *A* can be treated as a related concept of *A* and can be further expanded by describing other relations pertaining to *B*. However, the relations for *B* described therein will be "limited under the context of *A*". For example, if *A* is "organization" and *B* is the "facility" attribute of *A*, then underneath the node *B* we shall list those facilities one can normally find in an organization, whereas for the "facility" attribute of "hotel", we shall only list those existing facilities in hotel.

2.4 The Inference Engine of InfoMap

The kernel program can map a natural language sentence into a set of nodes and uses the edited knowledge to recognize the events in the user's sentences. Technically, InfoMap matches a natural language sentence to a collection of concept notes. There is a firing mechanism that finds nodes in InfoMap relevent to the input sentence. Suppose we want to find the event of the following sentence: "How do I invest in stocks?" and the interrogative word "how" can fire the word "method". Then along the path from "method" to "stock" the above sentence has fired the concepts "stock" and "invest". Thus, the above sentence will correspond to the path:

stock - *event* - invest - *attribute* - method

Given enough knowledge about the events related to the main concept, InfoMap can be used to parse Chinese sentences. Readers can refer to (Hsu et al. 2001) for a thorough description of InfoMap.

3. Automatic Domain Ontology Acquisition

To build an ontology for a new domain, we need to collect domain keywords and find the relationships among them. An acquisition process, SOAT, is designed that can construct a new ontology through domain corpus. Thus, with little human intervention, SOAT can build a prototype of the domain ontology.

As described in previous sections, InfoMap consists of two major relations among concepts, i.e., Taxonomic relations (category and synonym) and Non-taxonomic relations (attribute and event). We defined sentence templates, which consists of patterns of keywords and variables, to capture these relations.

3.1 Description of SOAT

Given the domain corpus with the POS tag, our SOAT can be described as follows.

Input: domain corpus with the POS tag
Output: domain ontology prototype
Steps:

1 Select a keyword (usually the name of the domain) in the corpus as the seed to form a potential root set *R*
2 Begin the following recursive process:
  2.1 Pick a keyword *A* as the root from *R*
  2.2 Find a new related keyword *B* of the root *A* by extraction rules and add it into the domain ontology according to the rules.
  2.3 If there is no more related keywords, remove *A* from *R*
  2.4 Put *B* into the potential root set
  2.5 Repeat step 2, until either *R* becomes empty or the total number of nodes generated exceeds a prescribed threshold.

We find that most of the domain keywords are not in the dictionary. So the traditional TF/IDF method would fail. Instead, we use the high frequency new words discovered by PAT-tree as the seeds. Ideally, SOAT can generate an domain ontology prototype automatically. However, the extraction rules need to be refined and updated by a human editor. The details of SOAT extraction rules are in Section 3.2.

3.2 The Extraction Rules of SOAT

The extraction rules in Tables 1, 2, 3 and 4, consists of a specific noun as the root, and the POS tags of the neighboring words. A rule is a linguistic template for finding keywords related to the root. The target of extraction is usually a word or a compound word, which has strong semantic links to the root. Our rules are especially effective in identifying essential compound words for a specific domain.

We use POS tags defined by CKIP (CKIP 1993), in which *Na* is the generic noun, *Nb* is the proper noun, and *Nc* is the toponym. Generally, an *Na* can be a subject or an object in a sentence, including concrete noun and abstract noun, such as “cloth”, “table”, “tax”, and “technology”. An *Nc* is the name of a place. Readers can refer to CKIP (CKIP 1993) for more information about the POS tag. In our experiment, we focus on *Na* and *Nc*, because the topics that we are interested in usually fall in these two categories. The extraction rules of finding categorical (taxonomy) relationships from a given *Na* (or *Nc*) are in Table 1 (and 3). The rules of finding attribute (non-taxonomy) relationships from a given *Na* (or *Nc*) are in Table 2 (and 4).

Table 1. Category extraction rules of an *Na* noun

Extraction rule	Extraction target	Example
A+Na（root）	A	信託（A）**股票（Na）**
Na+Na（root）	Na	水泥（Na）**股票（Na）**
Nb+Na（root）	Nb	三陽（Nb）**股票（Na）**
Nc+Na（root）	Nc	台泥（Nc）**股票（Na）**
Ncd+Na（root）	Ncd	
VH+Na（root）	VH	上市（VH）**股票（Na）**
Nc+Nc+Na（root）	Nc+Nc	華航（Nc）公司（Nc）**股票（Na）**
Na+Na+Na（root）	Na+Na	自營商（Na）庫存（Na）**股票（Na）**
VH+Na+Na（root）	VH+Na	公營（VH）事業（Na）**股票（Na）**

Table 2. Attribute extraction rules of an *Na* noun

Extraction rule	Extraction target	Example
Na（root）+Na	Na	**網路（Na）**主機（Na）
Na（root）+Nc	Nc	**網路（Na）**中心（Nc）
Na（root）+ DE +Na	Na	**網路（Na）**的(DE)連接埠（Na）

Table 3. Category extraction rules of an *Nc* noun

Extraction rule	Extraction target	Example
A+Nc（root）	A	縣立(A)**銀行(Nc Root)**
Na+Nc（root）	Na	政府(Na)**銀行(Nc Root)**
Nb+Nc（root）	Nb	花旗(Nb)**銀行(Nc Root)**
Nc+Nc（root）	Nc	台灣(Nc)**銀行(Nc Root)**
Ncd+Nc（root）	Ncd	
VH+Nc（root）	VH	民營（VH）**銀行（Nc**

		Root）
Na+Nb+Nc（root）	Na+Nb	英商（Na）柏克萊（Nb）**銀行（Nc Root）**
Nb+Na+Nc（root）	Nb+Na	世華（Nb）商業（Na）**銀行（Nc Root）**
Nb+VH+Nc（root）	Nb+VH	
Nc+A+Nc（root）	Nc+A	東京（Nc）共同（A）**銀行（Nc）**
Nc+FW+Nc（root）	Nc+FW	
Nc+Na+Nc（root）	Nc+Na	加拿大（Nc）皇家（Na）**銀行（Nc Root）**
Nc+Nb+Nc（root）	Nc+Nb	香港（Nc）匯豐（Nb）**銀行（Nc Root）**
Nc+VC+Nc（root）	Nc+VC	中國（Nc）建設（VC）**銀行（Nc Root）**
Nc+Nc+Na+Nc（root）	Nc+Nc+Na	中國（Nc）國際（Nc）商業（Na）**銀行（Nc Root）**
Nc+Nc+VC+Nc（root）	Nc+Nc+VC	中國（Nc）國際（Nc）開發（VC）**銀行（Nc Root）**

Table 4. Attribute extraction rules of an *Nc* noun

Extraction rule	Extraction target	Example
Nc（root）+Na	Na	**中央研究院（Nc）**院長（Na）
Nc（root）+Nc	Nc	**中央研究院（Nc）**停車場（Nc）
Nc（root）+Nc+Nc	Nc+Nc	**中央研究院（Nc）**語言所（Nc）語音實驗室（Nc）
Nc（root）+DE+Na	Na	**中央研究院（Nc）**的(DE)出版品（Na）

4. Discussion

Li and Thompson (1981) describe Mandarin Chinese as a Topic-prominent language in which the subject or the object is not as obvious as in other languages. Therefore, the highly precise shallow parsing result (Munoz et al. 1999) on NN and SV pairs in English is probably not applicable to Chinese.

4.1 The Experiment of Extraction Rate

To test the qualitative and quantitative performance of SOAT, we design two experiments. We construct three domain ontology prototypes for three different domains and corpora. Table 5 shows the result in which the frequently asked questions (FAQs) for stocks are taken from test sentences of the financial QA system. The university and bank corpora are collected from the CKIP corpus (CKIP 1995). We select sentences containing the keyword "University" or "Bank" as the domain corpora. The results in Table 5 show that SOAT can capture related keywords and the relationships among them from limited sentences very efficiently without using the frequency.

Table 5. The Extraction Rate in Different Domains

	Domains		
	Stock	**University**	**Bank**
Corpus	FAQ question	CKIP corpus	CKIP corpus
Sentences : ***S***	3385	3526	785
Extrated Nodes : ***N***	1791	2800	120
Extration Rate : ***N/S***	0.53	0.79	0.15

4.2 Results from Different Corpora

We select three different corpora from different information resources in the "network" domain. The first corpus is a collection of FAQ sentences about computer network. The second corpus is a collection of sentences containing the keyword "network" from the CKIP corpus. The third corpus is the collection of sentences from Windows 2000 online help document. To reduce the cost of human verification, we limit the size of corpus to 275 sentences. The result in Table 6 shows that there is a trade-off between extraction rate and the accuracy rate.

Table 6. The extraction and accuracy rate of three corpora in the same domain

	Network Domain		
Corpus	FAQs	CKIP corpus	Online help documents
Sentences : ***S***	275	275	275
Extracted Nodes : ***N***	25	180	73
Extraction Rate : ***N/S***	0.09	0.65	0.27
Human verified: ***H***	19	25	45
Accuracy rate : ***H/N***	0.76	0.14	0.62

4.3 The Advantage of a Semi-Automatic Domain Ontology Editor for QA System

SOAT can help in QA system ontology editing. In our experience, a trained knowledgeable editor can compile about 100 FAQs into our ontology manually per day. On the other hand, with the help of SOAT, a knowledgeable editor can edit on the average 4 categories, 25 attributes and 42 activities that SOAT extracted.

The quantity is estimated on 4*(25+42)=268 FAQ query concepts at least. Thus, the productivity of using SOAT is approximated 268% times. It is obvious that SOAT can help reducing the cost of building a new domain ontology.

5. Conclusion

We present a semi-automatic process of domain ontology acquisition from domain corpus. The ontology schema we used is general enough for different applications and specific enough for the task of understanding the Chinese natural language. The main objective of the research is to extract useful relationships from domain articles to construct domain ontology prototypes in a semi-automatic fashion. The SOAT extraction rules we developed can identify keywords with strong semantic links, especially those compound words in the domain.

We have discussed how to extract related NN pairs in Section 3 for SOAT. However, the extraction rules for NN pairs do not apply for NV pairs. In the future we shall follow the approach in (Tsai et al. 2002) to extract the relationships between nouns and its related verbs.

The main restriction of SOAT is that the quality of the corpus must be very high, namely, the sentences are accurate and abundant enough to include most of the important relationships to be extracted.

References

Abney, S.P. (1991), Parsing by chunks. In Berwick, R.C., Abney, S.P. and Tenny, C. (ed.), Principle-based parsing: Computation and Psycholinguistics, pp. 257-278. Kluwer, Dordrecht.

Bredenkamp, A., Crysmann, B., and Petrea, M. (2000), Looking for Errors: A declarative formalism for resource-adaptive language checking, Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.

CKIP (1993), Chinese Part-of-speech analysis, Technical Report 93-05, Academia Sinica, Taipei.

CKIP (1995), A Description to the Sinica Corpus, Technical Report 95-02, Academia Sinica, Taipei.

Flett, A. and Brown, M. (2001), Enterprise-standard Ontology Environments for Intelligent E-Business, Proceedings of IJCAI-01 Workshop on E-Business & the Intelligent Web, Seattle, USA.

Gruber, T.R. (1993), A translation approach to portable ontologies. Knowledge Acquisition, 5(2), pp. 199-220, 1993.

Guarino, N. (1998), Formal Ontology and Information Systems, Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, pp. 3-15. IOS Press.

Hearst, M.A. (1992), Automatic acquisition of hyponyms from large text corpora. In COLING-92, pp. 539-545.

Hsu, W.L., Wu, S.H. and Chen, Y.S. (2001), Event Identification Based On The Information Map - InfoMap, in symposium NLPKE of the IEEE SMC Conference, Tucson Arizona, USA.

Li, C.N. and S.A. Thompson (1981), Mandarin Chinese: a functional reference grammar, University of California press.

Maedche, A. and Staab, S. (2000), Discovering Conceptual Relations from Text. In: Horn, W. (ed.): ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam.

Munoz, M., Punyakanok, V., Roth, D., Zimak, D. (1999), A Learning Approach to Shallow Parsing, Proceedings of EMNLP-WVLC'99.

Noy, N.F. and McGuinness D.L. (2001), Ontology Development 101: A Guide to Creating Your First Ontology, SMI technical report SMI-2001-0880, Stanford Medical Informatics.

Tsai, J. L, Hsu, W.L. and Su, J.W. (2002), Word sense disambiguation and sense-based NV event-frame identifier. Computational Linguistics and Chinese Language Processing, 7(1), pp. 1-18.

Wu, S.H., Day, M.Y., Tsai, T.H. and Hsu, W.L. (2002), FAQ-centered Organizational Memory, in Matta, N. and Dieng-Kuntz, R. (ed.), Knowledge Management and Organizational Memories, Kluwer Academic Publishers.