Proceedings of the Second Conference for Natural Language Processing
(KONVENS-94), Vienna, Austria, September 28-30, 1994, pp. 407-410.

Acquisition of Subcategorization Frames from Large Scale Texts

Kuang-hua Chen and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
Internet: hh_chen@csie.ntu.edu.tw

Abstract. Subcategorization frames are useful for many applications. Due to many ambiguities, to extract them is not straightforward. In this paper, a probabilistic chunker is used to determine the plausible phrase boundaries and a finite state mechanism, SUBCAT-TRACTOR, is proposed to extract 23 subcategorization frames. In order to get rid of the problems introduced by compound nouns, a noun-phrase extractor is applied. In addition, two extra rules are presented to capture the movement phenomena.

Abstakt. Unterkategorierungs-Rahmen sind vielseitig verwendbar. Diese Rahmen können jedoch nicht direkt ausgewählt werden. In diesem Aufsatz geht es um einen "Probabilistic Chunker", mit dem man sinnvolle Satz-Grenzen bestimmen kann und um einen "Subcat-Tractor", mit dem 23 Unterkategorierungs-Rahmen ausgewählt werden können. Um die probleme zu lösen, die derch zusammengesetzte Nomen entstehen, wird einen "Noun-Phrase Extractor" verwendet. Zusätzlich werden zwei Regeln, bezüglich der Bewegungs-Erscheinung vorgestellt.

1. Introduction

Manually compiling lexical knowledge, e.g., lexicon, subcategorization frames, *etc*., is time-consuming and suffers from inconsistency. Automatic compilation of these data not only relieves the shortcomings, but also reflects the up-to-date usages. Some researchers apply shallow or partial parsers to acquiring specific patterns from texts [1,2]. These show that it is not necessary to fully parse the texts for some applications. This paper will combine the statistical approach and linguistic theory to extract subcategorization frames from text corpora. A probabilistic chunker [3] is used to segment the texts into chunks and each chunk is assigned a head. Then, we further collect the noun phrases by NOUN-TRACTOR [4] to avoid the problems introduced by compound nouns. Finally, a finite state mechanism based on linguistic theory is used to acquire these subcategorization frames.

2. Acquisitions of Verb Subcategorization Frames

Some dictionaries offer verb subcategorization frames, e.g., Oxford Advanced Learner's Dictionary (OALD) [5] classifies 32 different verb frames to describe the usages of verbs. However, OALD only offers the possibilities of the usages of verb subcategorization frames. On the one hand, how many probabilities these structures have is not clear. On the other hand, to update a dictionary timely is expensive, and it is easy to induce inconsistency.

A approach to acquisition of subcategorization frames is proposed in this paper. To extract subcategorization frames, each text is segmented into plausible phrasal chunks. Then, we further collect the noun phrases to avoid the problems introduced by compound nouns. In addition, rules are also presented to capture the movement phenomena. These issues are not seen in the other literatures. Finally, a finite state mechanism is utilized to extract the verb subcategorization frames. Since some frames defined by OALD cannot be distinguished from each other by the contexts, we define the hypertypes for these frames. These hypertypes are shown in Table 1.

Table 1. Hypertypes

Hypertypes	Types	Hypertypes	Types	Hypertypes	Types	Hypertypes	Types
Vnpr	Dnpr, Tnpr	Vt	It, Tt	Tf	Tf	Dnf	Dnf
Vn	Ln, Tn	La	La	Tw	Tw	Dprf	Dprf
Vnt	Cnt, Dnt, Tnt	I	I	Tg	Tg	Dnw	Dnw
Vni	Cni, Tni	Ipr	Ipr	Tsg	Tsg	Dprw	Dprw
Vng	Cng, Tng	Ip	Ip	Cna	Cna	Dprt	Dprt
Vnn	Cnn, Dnn	Tnp	Tnp	Cnn/a	Cnn/a		

That is to say, Ln and Tn will be recognized as Vn and the real type is decided by the verb itself and the type. In addition, the subcategorization frame In/pr is regarded as In and Ipr. Therefore, there are 23 verb frames in total in our experiments. Figure 1 shows the finite state mechanism, SUBCAT-TRACTOR, and double circle denotes the 23 frames. The capital symbols on the arcs in Figure 1 are tags defined by LOB corpus [6]. The "as" denotes word *as*, "wh*" denotes the wh-word, and "that" denotes word *that*. * means the wildcard. In addition, two rules are proposed to capture the movements by relativization and passivization.

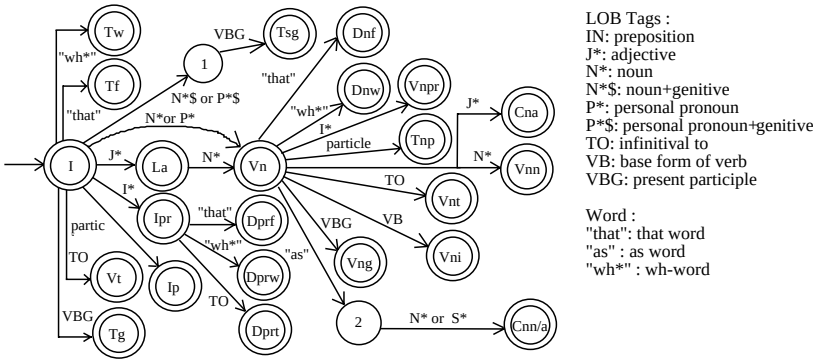

Figure 1. Finite State Mechanism for Acquisition of Subcategorization Frames

3. Experiments and Comparisons

We uses SUSANNE Corpus [7] as the test corpus. Three experiments are conducted to compare the impacts of noun phrases and special rules. Firstly, a baseline method uses the SUBCAT-TRACTOR to process the chunked texts and retrieves the subcategorization frames. The second method (Enhanced I) is the baseline method

enhanced by NOUN-TRACTOR. The third method (Enhanced II) is the Enhanced I method further enhanced by two extra rules. These two rules deal with passivization and relativization. Move-a [8] postulates that a constituent is moved to the landing site, and a trace is left at the empty site. In general, the moved constituent is an NP, and the object position is a possible empty site. Thus, the effect of movement transformation must be considered when subcategorization frames are formulated. It is not easy to tell out adjunct and argument, so that our system always regards the prepositional phrases as argument. The experimental results are shown in Table 2.

Table 2. Comparisons of Different Methods

		Cna	Cnn/a	Dnf	Dnw	Dprf	Dprt	Dprw	I	Ip	Ipr	La	Tf
Baseline	Verb	74	11	80	14	24	13	11	987	264	737	110	119
	Freq.	103	11	167	23	24	13	11	4397	562	2255	914	443
Enhanced I	Verb	103	10	87	72	14	9	4	1130	46	981	123	130
	Freq.	173	10	150	122	14	9	4	4714	57	3212	953	426
Enhanced II	Verb	159	10	105	75	14	8	4	1017	27	719	105	120
	Freq.	264	10	183	126	14	8	4	4212	33	2447	924	388

		Tg	Tnp	Tsg	Tw	Vng	Vni	Vnn	Vn	Vnt	Vnpr	Vr	
Baseline	Verb	50	142	17	47	90	37	278	847	95	455	133	
	Freq.	95	296	23	121	186	143	723	3694	248	1125	596	
Enhanced I	Verb	74	32	11	50	87	46	106	937	99	535	143	
	Freq.	112	35	11	109	143	140	222	3362	225	1460	563	
Enhanced II	Verb	66	53	11	45	91	45	141	1019	133	739	101	
	Freq.	103	60	11	103	152	139	287	3782	321	2141	469	

The frequency of Vnn generated by the second method (Enhanced I) is significantly less than that of Vnn generated by the baseline method. Many Ipr patterns are misregarded as Ip in baseline method. That results from the utilization of NOUN-TRACTOR. (1) is an example.

(1) ... gave Norm Van Brocklin permission ...
(SUSANNE A14:0100i-A14:0110a)

Comparing Enhanced I and Enhanced II methods, the frequencies of I patterns, Ip patterns and Ipr patterns are reduced in the Enhanced II method. That is to say, we capture many occurrences of movements. For example, the surface subject is moved from the object of *give* in (2) by passivization.

(2) The impression has nevertheless been given during these three days, ...
(SUSANNE A04:0400d-A04:0410f)
... he would be given the right to pick Mr. Gerosa's successor, ...
(SUSANNE A07:0360b-A07:0360p)

Another problem is introduced by the usages of phrasal verb. This is shown in (3).

(3) But it might give way shortly to another vexing issue ...
(SUSANNE A08:1220c-A08:1220m)

The occurrence of verb *give* in (3) is regarded as Vnpr. Some other errors are discussed in the following. We first check whether the usage of "La" of verb *find* is correct or not in (4).

(4) However, there is always the possibility that chance will make
demands the dancers *find impossible* to execute.
(SUSANNE G09:1740k-G09:1760d)

Label the usage of *find* with "La" is not correct. It should be "Vat" (verb + adjective + to-infinitive) in the convention of OALD. However, the usage does not appear in the list of subcategorization frames of OALD. Unavoidably, it is taken as "La" by our system. The "Vng" usage of *find* is not listed in OALD, but we may find one in Longman Dictionary of Contemporary English (LDOCE) [9].

(5) They found the lost child hiding in the cave. [9:381]

This example shows that dictionaries are not complete, that is, some usages in running texts may not appear in the dictionaries.

4. Concluding Remarks

Many approaches are proposed to build verb subcategorization frames in dictionaries automatically. The statistics-based approach and linguistic theory are integrated to extract subcategorization frames in the paper. These two kinds of methods are complementary. Statistics-based approach is robust. It provides simple language models to analyze unrestricted texts. However, it may need large completely-annotated corpus to treat complex linguistic phenomena. Linguistic theory gives such a supplement. Well-formed patterns can be explained properly by universal principles, so that they can be formulated in terms of rules easily. The experimental results show that the integrated mechanisms are useful for further researches on large volume of real texts.

Acknowledgement

We are grateful to Dr. Geoffrey Sampson for his kindly providing SUSANNE Corpus and the details of tag set to us. This research was supported in part by National Science Council, Taipei, Taiwan, R.O.C. under contract NSC83-0408-E-002-019.

References

[1] K.W. Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proceedings of Second Conference on Applied Natural Language Processing*, 1988, pp. 136-143.

[2] M. Brent, "Automatic Acquisition of Subcategorization Frame from Untagged Text," *Proceedings of the 29th Annual Meeting of ACL*, 1991, pp. 209-214.

[3] K.-H. Chen and H.-H. Chen, "A Probabilistic Chunker," *Proceedings of the 6th ROCLING, 1993*, pp. 99-117.

[4] K.-H. Chen and H.-H. Chen, "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation," *Proceedings of the 32nd Annual Meeting of ACL*, New Mexico, 1994, pp. 234-241.

[5] A.S. Hornby, *Oxford Advanced Learner's Dictionary*, Oxford University Press, 1989.

[6] S. Johansson, *The Tagged LOB Corpus: Users' Manual*, Bergen: Norwegian Computing Centre for the Humanities, 1986.

[7] G. Sampson, "The Susanne Corpus," *ICAME Journal*, No. 17, 1993, pp. 125-127.

[8] P. Sells, *Lectures on Contemporary Syntactic Theories*, Lecture Notes, No. 3, CSLI, 1985.

[9] R. Quirk, *Longman Dictionary of Contemporary English*, Longman Group UK Limited, 1987.