A Framework of Automatic Case Frame Construction From a Raw Corpus

Gongye Jin Daisuke Kawahara Sadao Kurohashi
Graduate School of Informatics, Kyoto University

jin@nlp.ist.i.kyoto-u.ac.jp,{dk,kuro}@i.kyoto-u.ac.jp

1 Introduction

In order to let computers understand text, identifying various types of relations in the text is a necessary step. In natural language processing (NLP), case frames are a very essential knowledge base which represents the relations between a predicate and its arguments. Case frames can support various types of NLP applications, such as parsing, machine translation, recognizing textual entailment and paraphrasing. For instance, in the classical example of parsing, "saw a girl with a telescope," there is an ambiguity problem of which argument the prepositional phrase, "with a telescope," is modifying. It would be easy to judge that the prepositional phrase belongs to the verb if knowledge of a case frame "see someone/something with telescope," is available.

For NLP applications in multiple languages and multilingual applications, compilation of large-scale case frames in these languages is important. Manually compiling this kind of knowledge in multiple languages would be too costly and would be limited by low coverage. In this paper, we propose a framework for automatically constructing case frames in multiple languages. In this framework, we extract reliable predicate-argument structures from large corpora by small sets of linguistic rules specific to each language, and apply clustering to produce case frames. Even though the grammars in different languages are quite dissimilar, we show that extracting predicate-argument structures to build case frames is achieved by one common way. In practice, we construct case frames of Chinese and English.

2 Related Work

In early research, subcategorization frames were proposed and automatically constructed to represent the relations between verbs and other syntactic arguments in text (Brent, 1993). However, subcategorization frames do not have sufficient capacity for indicating semantic relations in text.

Japanese case frames have been automatically constructed by exploiting case-marking post-positions (Kawahara and Kurohashi, 2006). However, building case frames in other languages which do not have case markers is still a challenging task. As for English, a method for acquiring reliable predicate-argument structures from a raw corpus has been proposed (Kawahara and Kurohashi, 2010), but it is not for compiling case frames.

3 Framework for Case Frame Construction

3.1 Overview of Construction Framework

We construct case frames for each predicate. Case frames of one certain predicate are separated according to the predicate's usages. In Table 1, we show two case frames of the English verb *run*. To automatically acquire case frames, we first extract highly-reliable predicate-argument structures form a raw corpus. Then, we subsequently apply semantic clustering to distinguish the different usages of each predicate. Our construction method includes the following steps: preprocessing of raw texts, parsing, filtering out unreliable parses by language-specific rules, extracting predicate-argument structures and semantic clustering.

3.2 Analysis to Raw Corpus

Some languages such as Chinese and Japanese do not delimit words by white-space. For these languages, word segmentation is needed as a pre-process. To word sequence, we apply part of speach (POS) tagging in order to assign a syntactic category to words. We also apply chunking process to form a constituent such as verb phrases (VP) and noun phrases (NP).

3.3 Filtering Automatic Parses

There always exist sentences that are incomplete in grammar or syntax, and could not be used to cor-

verb	case slot	instance
run(1)	sbj	user:107 computer:79 ...
	obj	server:6082 programme:5085
	time	year:40 week:16 today:4
	pp:without	statement:730
	pp:on	network:76 <num>:51 ...
	...	...
run(2)	sbj	i:8695 it:6051 he:3672
	pp:via	list:1968
	pp:across	them:1976
	pp:into	problem:926 trouble:627
	...	...
...		

Table 1: Examples of English Case Frames

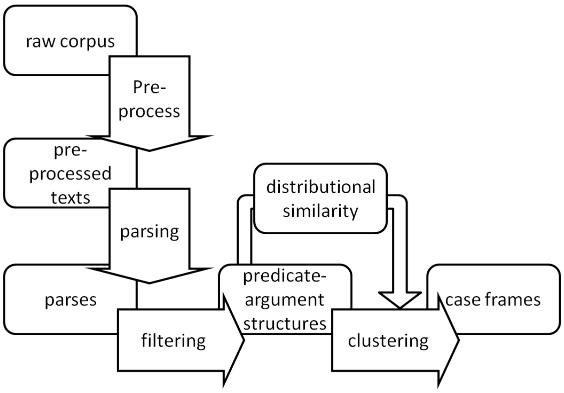

Figure 1: Overview of Construction

rectly provide the usages of predicates. Similarly, some parts of chunks with complicated syntax also make the complete sentence unreliable. We define small sets of language-specific linguistic rules to filter out unreliable parses. For example, we filter the sentences which satisfy the following conditions both for English and Chinese:

- a sentence which ends with a question mark
- a sentence which includes sign

We also prune unreliable parts of a sentence such as:

- all the chunks after comma in English
- all the chunks before the second VP from bottom in Chinese

By applying this set of rules, we can avoid some complex cases which contain several verbs, especially some clauses led by *WH* which involve more complex dependency relations. Unlike English, Chinese nearly does not have any marker words for clause such as where or which etc. In the sentence such as 我*(I)* 希望*(hope)* 他*(he)* 早日*(soon)* 康复*(recover)*, we only maintain the part that surround the last verb. Therefore, this sentence will be prunned as 他*(he)* 早日*(soon)* 康复*(recover)*.

3.4 Extracting Predicate-argument Structures

From the reliable parses, predicate-argument structures are extracted in a straightforward way. We convert the *VP* to *pred*, and the *NP* preceding the predicate is converted to *sbj*. An *NP* following the predicate is converted to *obj*. In *BA* phrase of Chinese, we change *BA* and *LB* to *ba*[1]. A *PP* which is combined with *NP* is converted to a prepositional phrase. For example, one predicate-argument structure can be written as *sbj:I pred:put obj:pen pp:on:desk*.

3.5 Clustering Predicate-argument Structures

After the extraction of predicate-argument structures, we cluster the predicate-argument structures into their usages. For each predicate-argument structure, we firstly choose one important argument which can mostly indicate the predicate's meaning and is for most of the time the object. We call it "key phrase". We initially group the instances with the same key phrase to initial clusters. To implement further clustering of all the initial clusters, we utilize the method of similarity calculation similar to Japanese (Kawahara and Kurohashi, 2006), which considers two aspects of initial clusters: the similarity of case slot patterns and the similarity between the words in the same position of case slot. To calculate the similarity between two words, we use distributional similarity, which is based on the hypothesis that words with similar semantic features always share the similar contexts (Lin, 1998). This distributional similarity is calculated based on the extracted predicate-argument structures.

4 Experiments

4.1 Construction of English and Chinese Case Frames

For English case frames we made use of 200 million sentences extracted from the Web and for Chinese we utilized Center for Chinese Linguistic (CCL) corpus which contain 10 million sentences for construction. We used the MSTparser (McDonald et al. 2006) for parsing.The file wsj_02-21 in PennTreebank (PTB) are used as training data to train the parser for English. In ChineseTreebank (CTB), we use file chtb_001-270, chtb_400-931, chtb_1001-1151 as training data, We extracted predicate-argument structrues for about 14,000 English predicates and 10,000 Chinese predicates.

[1] As for the BA phrase, it is a special case of grammar in Chinese which makes the object behind the verb and change the original phrase order of SVO to SOV which is similar to Japanese, thus the meaning is almost the same.

language	precision	recall
English	0.977	0.357
Chinese	0.982	0.337

Table 2: Automatic Evaluation of Predicate-argument Structures

4.2 Evaluation

We evaluate the acquired case frames in two ways. First, we automatically judge how reliable predicate-argument structures are produced by the proposed framework. Second, to see the effectiveness of clustering, we manully evaluate the clustered case frames and compare them with subcategorization frames which are constructed as a baseline.

4.2.1 Automatic Evaluation of Predicate-argument Structures

In order to evaluate the filtering rules that we used to extract highly-reliable predicate-argument structures, we automatically evaluated them using PT-B for English and CTB 5.0 for Chinese. We use file wsj_23 to test for English and file chtb_271-300 in CTB as testing data for Chinese. First, we applied our framework to the raw texts in the treebank, and extracted predicate-argument structures. In each predicate-argument structure we extracted, the dependency relation is defined as: every argument depends on the predicate. Then, we use the existing dependency annotation in each treebank to extract gold standard dependency pairs. We calculate the precision as the percentage of correct dependency pairs among acquired pairs. We calculate the recall as percentage of correct dependency pairs among the total dependency pairs in the treebank.

As we can see in Table 2, for both Chinese and English we achieved a precision of over 97% and a recall of around 30%. From our error analysis, most of the incorrect dependency pairs is due to dependency parsing errors. For example, in the Chinese sentence “撞击(impact)事件(event)”, the noun “事件” is always parsed as an arugment of the verb “撞击”. However, there is an omitted character “的” between these two words and the verb “撞击” is actually a modifier of the noun “事件”. This problem can be solved by using the feedback of constructed case frames in the future work.

4.2.2 Manual Evaluation of Case Frames

We built subcategorization frames from the same corpora for each predicate for comparison. For both subcategorization frames and case frames, we conducted two types of evaluation: slot-based and frame-based. For the slot-based evaluation, We manually judged whether each case slot is good by the criterion that 80% of the instances in the case slot are semantically similar. We then calculated the accuracy of both kind of frames by the percentage of good case slots.

As the above evaluation is based on case slots, we also conducted a frame-based evaluation. For each case frame we built, it can be seen as a good frame only if it satisfies the following two conditions. First, above 80% of its key phrases are semantically similar. Second, the key phrases in the frame must be semantically independent with any other existing frames. That is, one case frame must not be similar with any others. The accuracy is calculated as the percentage of good frames. In subcategorization frames, we simple choos the noun directly after or before the predicate to be the key phrase.

The evaluation results for seven frequent English and Chinese verbs are shown in Tables 3, Table 4, Table 5 and Table 6. As we can see, the clustering method in the construction of case frames merged most similar instances in each case slot, so the total number of case frames is much smaller than subcategorization frames. And case frames outperformed the subcategorization frames obviously, because subacategorization frames is only clustered by syntactic patterns.

4.3 Discussion

Many linguistic variations cause parsing errors and lead to the decrease of accuracy in the acquisition of predicate-argument structures. In Chinese, for example, the character “的” always causes not only syntactic but also semantic ambiguity. Also, omission is one common feature of natural languages. For instance, many clause markers like *which, where* or *that* are missing frequently. This leads to wrong results in our system. For example, in the sentence *I heard the machine exploded*, our method incorrectly make *the machine* the direct object of *heard*. This kind of problem is to be solved by the feedback from constructed case frames in the future research.

5 Conclusion

In this paper, we propose a framework for automatically constructing case frames for multiple languages. Our framework successfully extracted high-reliable predicate-argument structures from corpora and well clustered instances with similar semantic features to produce the final case frames. We plan to improve the case frames for each language by feedbacking information in the case frames, We also plan to use a larger-scale corpus such as the Web corpus of Chinese. After building large-scale multilingual case frames, we have a plan to use them in machine translation to assist alignment.

predicate	subcat frames	case frames
visit	0.44 (80/181)	0.67 (16/24)
run	0.20 (40/201)	0.50 (36/72)
begin	0.18 (33/179)	0.52 (42/80)
believe	0.39 (27/70)	0.55 (20/36)
ask	0.28 (43/156)	0.52 (14/27)
find	0.17 (55/332)	0.53 (9/17)
add	0.22 (46/208)	0.52 (14/27)
total	0.27	0.54

Table 3: Slot-based Evaluation of English Case Frames

predicate	subcat frames	case frames
产生	0.37 (13/35)	0.54 (15/28)
发表	0.35 (59/111)	0.59 (32/54)
发展	0.26 (40/154)	0.54 (26/48)
贡献	0.22 (4/18)	0.71 (5/7)
进入	0.25 (28/111)	0.61 (9/17)
开展	0.28 (31/110)	0.51 (31/60)
提出	0.29 (40/141)	0.57 (60/105)
total	0.29	0.58

Table 4: Slot-based Evaluation of Chinese Case Frames

predicate	subcat frames	case frames
visit	0.24 (8/33)	0.75 (3/4)
run	0.05 (2/38)	0.67 (4/6)
begin	0.15 (5/34)	0.73 (8/11)
believe	0.28 (5/18)	0.80 (3/5)
ask	0.39 (13/33)	1.00 (2/2)
find	0.30 (18/60)	0.75 (6/8)
add	0.18 (7/39)	0.78 (7/9)
total	0.23	0.78

Table 5: Frame-based Evaluation of English Case Frames

predicate	subcat frames	case frames
产生	0.11 (4/36)	0.70 (7/10)
发表	0.21 (6/28)	0.50 (8/16)
发展	0.09 (4/43)	0.71 (15/21)
贡献	0.25 (2/8)	0.67 (2/3)
进入	0.10 (3/29)	0.71 (15/21)
开展	0.10 (3/31)	0.73 (11/15)
提出	0.21 (8/38)	0.67 (13/20)
total	0.15	0.66

Table 6: Frame-based Evaluation of Chinese Case Frames

References

[1] Michael R. Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19, pages 243–262

[2] Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex Syntax: building a computational lexicon. In *Proceedings of the 15th conference on Computational linguistics*, pages 268–272.

[3] Daisuke Kawahara and Sadao Kurohashi. 2006. Caseframe Compilation from the Web using High- Performance Computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1344–1347.

[4] Daisuke Kawahara and Sadao Kurohashi. 2010. Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1389–1393.

[5] Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. In *Proceedings of ACL-IJCNLP 2009*, pages 513–521.

[6] Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional Inference with the Easiest-First S-trategy for Tagging Sequence Data. In *Proceedings of HLT/EMNLP*, pages 562–568.

[7] Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two stage discriminative parser. In *Proceedings of CoNLL*, pages 216–220.

[8] D. Lin. 1998 An information-theoretic definition of similarity. In *Proceeding of the Fifteenth International Joint Conference on Machine Learning*, page 296–304.