Resolving Ambiguities of Chinese Conjunctive Structures by Divide-and-conquer Approaches

Duen-Chi Yang, Yu-Ming Hsieh, Keh-Jiann Chen
Institute of Information Science, Academia Sinica, Taipei
{ydc, morris, kchen}@iis.sinica.edu.tw

Abstract

This paper presents a method to enhance a Chinese parser in parsing conjunctive structures. Long conjunctive structures cause long-distance dependencies and tremendous syntactic ambiguities. Pure syntactic approaches hardly can determine boundaries of conjunctive phrases properly. In this paper, we propose a divide-and-conquer approach which overcomes the difficulty of data-sparseness of the training data and uses both syntactic symmetry and semantic reasonableness to evaluate ambiguous conjunctive structures. In comparing with the performances of the PCFG parser without using the divide-and-conquer approach, the precision of the conjunctive boundary detection is improved from 53.47% to 83.17%, and the bracketing f-score of sentences with conjunctive structures is raised up about 11 %.

1 Introduction

Parsing a sentence with long conjunctive structure is difficult, since it is inadequate for a context-free grammar to represent context-sensitive-like coordination structures, such as “a b c… and a’ b’ c’… ”. It causes long-distance dependencies and tremendous syntactic ambiguities (a large number of alternatives). Pure syntactic approaches cannot determine boundaries of conjunctive phrases properly. It is obvious that both syntactic and semantic information are necessary for resolving ambiguous boundaries of conjunctive structures.

Some analysis methods of the detection of conjunctive structures have been studied for a while. Despite of using different resources and tools, these methods mainly make use of the similarity of words or word categories on both sides of conjunctive structure (Agarwal et al., 1992; Kurohashi et al., 1994; Delden, 2002; Steiner 2003). They assumed that two sides of conjuncts should have similar syntactic and semantic structures. Some papers also suggest that certain key word patterns can be used to decide the boundaries (Wu 2003). Agarwal et al. (1992) used a semantic tagger and a syntactic chunker to label syntactic and semantic chunks. And then they defined multi-level (category to category or semantic type to semantic type) similarity matching to find the structure boundaries. Delden (2002) included semantic analysis by applying WordNet (Miller 1993) information. These presented methods used similarity measures heuristically according to the property of the languages. However detecting conjunctive boundaries with a similar method in Chinese may meet some problems, since a Chinese word may play different syntactic functions without inflection. It results that syntactic symmetry is not enough to resolve ambiguities of conjunctive structures and semantic reasonableness is hard to be evaluated. Therefore we propose a divide-and-conquer approach which takes the advantage of using structure information of partial sentences located at both sides of conjunction. Furthermore we believe that simple cases can be solved by simple methods which are efficient and only complex cases require deep syntactic and semantic analysis. Therefore we develop an algorithm to discriminate simple cases and complex cases first. We then use a sophisticated algorithm to handle complex cases only.

For simple cases, we use conventional pattern matching approach to speedup process. For complex conjunctive structures, we propose a divide-and-conquer approach to resolve the problem. An input sentence with complex conjunctive structure is first divided into two parts, one to the left of the conjunctive and one to the right, and then parsed independently to detect possible candidates of two conjuncts. The particular property of complex conjunctive structures of Chinese language allows us to parse and to produce syntactic structures of two partial sentences, since according to our observations and experiments the syntactic structures of partial sentences at either side of a complex conjunctive construction are grammatical most of the times. Figure 1 shows an instance. The parsing results not only reduce the possible ambiguous boundaries but also provide global structural information for checking the properness of both sides of conjunctive structure. Another important point worth mentioning is that since the size of available Treebank is small, a two-stage approach is proposed to resolve the data sparseness problems in evaluating syntactic symmetry and semantic reasonableness. At the first stage, a Conditional Random Fields model is trained and used to generate a set of candidate boundaries. At the second stage, a word-association model is trained from a gigaword corpus to evaluate the semantic properness of candidates. The proposed divide-and-conquer algorithm avoids parsing full complex conjunctive structures and handles conjunctive structures with deep structural and semantic analysis.

The extraction method for context-dependent rules is described in Section 2 and detail of the divide-and-conquer approach is stated in Section 3. In Section 4, we introduce our experimental environment and show the results of our experiment. We also make some discussions about our observations in Section 4. Finally, we offer our conclusion and future work in Section 5.

2 Boundary Detection for Simple Conjunctive Phrases

The aim of this phase of approach is to determine if simple conjunctive phrases exist in input sentences and then identify their boundaries by matching context-dependent rules. To derive a set of context-dependent rules for conjunctive phrases, a naïve approach is to extract all conjunctive patterns with their contextual constraints from Treebank. However such a set of extracted rules suffers a low coverage rate, since limited size of training data causes zero frequency of long n-gram PoS patterns.

2.1 Rule extraction and generalization

Agarwal et al., (1992), Kurohashi et al., (1994), and Delden (2002) had shown that the properties of likeness and symmetry in both syntactic types and lengths for example, exist in most conjunctive cases. Hence we use both properties as the conditions in deciding boundaries of conjunctive phrases. When we observe Sinica Treebank (Chen et al., 2003), we also find that this property is more obvious in simple conjunctive cases than in complex cases.

First, we use a simple algorithm to detect the boundaries of completely symmetric conjunctive phrases. If PoS patterns of “A B C and A B C” or “A B and A B” occurred in the input sentence, we consider patterns of such structures are legitimate conjunctive structures regardless whether the PoS sequences “A B C and A B C” or “A B and A B” ever occurred in the Treebank. For other cases we use context-dependent rule patterns to determine boundaries of conjunctive structures.

Statistical context-dependent PoS-based rule patterns are extracted automatically from Sinica Treebank. Each rule contains the PoS pattern of a conjunctive phrase and its left/right contextual constraints. The occurrence frequency of the rule and its correct identification rate are also associated. e.g. [VC] (Na Caa Nc) [DE][1] ; 12; 11

This rule says that PoS sequence Na Caa Nc forms a conjunctive phrase when its left context is a VC and its right context is a DE. Such pattern occurred 12 times in the training corpus and 11 out of 12 times (Na Caa Nc) are correct conjunctive phrases.

Context-dependent rule patterns are generated and generalized by the following procedure.

Rule Generation and Generalization

For each conjunctive structure in the Treebank, we consider a window pattern of at most 9 words. This pattern contains conjunction in the center and at most 4 words at each side of the conjunction. The PoS sequence of these 9 words forms a context-dependent rule. For instance, the conjunctive structure shown in Figure 1 will generate the pattern (1).

(1) [Vc DM] (VH Na Caa Neu Na) [DE Na]

The long pattern has low applicability and hardly

[1] Caa is a PoS for coordinate conjunction. Na is a common noun; Nc denotes place noun, and Vc is a transitive verb. DE denotes the relativizer ‘的’.

can evaluate its precision. Therefore a rule generalization process is applied. Two kinds of generalizations are available. One is reducing the length of contextual constrains and the other is to reduce a fine-grained PoS constraint to a coarse-grained PoS. Some instances, shown in (2), are the generalized patterns of (1).

(2) [DM] (VH Na Caa Neu Na) [DE];1;1
(VH Na Caa Neu Na); 10; 5
[DM] (V N Caa N N) [DE]; 3; 2

Then the applicability and precision of rules higher than threshold values will be selected. The threshold values for the rule selection are determined by testing results on the development data.

3 Resolution of Complex Conjunctive Structures

Complex structures are cases whose boundaries can not be identified by the pattern matching at phase-1. We propose a divide-and-conquer approach to resolve the problem. An input sentence with complex conjunctive structure was first divided into two parts with each part containing one of the conjuncts and then parsed independently to produce their syntactic structures for detecting possible boundaries of two conjuncts. Then ambiguous candidate structures are generated and the best conjunctive structure is selected by evaluating syntactic symmetry and semantic reasonableness of the candidates. Since the two parts of the partial sentences are simple without conjunctive structure and normally grammatical[2], hence they can be easily parsed by a PCFG parser.

Here we illustrate the divide-and-conquer algorithm by the following example. For instance, the example shown in Figure 1 has complex conjunctive structure and it was first split into two parts (1a) and (1b) at conjunction marker "、".

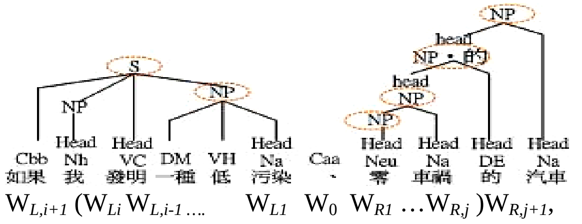
(1a) 如果 *if* (Cbb) 我 *I* (Nh) 發明 *invent* (VC) 一種 *a kind* (DM) 低 *low* (VH) 汙染 *pollution* (Na)

(1b) 零 *null* (Neu) 車禍 *accident* (Na) 的(DE) 汽車 *car* (Na)

The two parts of partial sentences are then parsed to produce their syntactic structures as shown in Figure 1. Then a CRF model trained from Sinica Treebank for checking syntactic symmetry was derived to pick the top-N candidates according to the syntactic information of both sides of partial sentences. Then at the second stage, a semantic evaluation model is proposed to select the best candidate. The detail of the semantic evaluation model is described in the section 3.2. The reason for using a two-stage approach is that the size of the Treebank is limited, but the semantic evaluation model requires the values of association strengths between words. The current Treebank cannot provide enough coverage and reliable values of word-association strengths.

3.1 Derive and evaluate possible candidates

CRF is a well-known probabilistic framework for segmenting and labeling sequence data (Lafferty, et al. 2001). In our experiments, we regard the problem of boundary detection as a chunking-like problem (Lee et al., 2005). Due to this reason, we use CRF model to generate candidates and their ranks. The features used in CRF model included some global syntactic information, such as syntactic category of a partial structure and its phrasal head. Such global syntactic information is crucial for the success of boundary detection and is not available if without the step of parsing process.



Some example feature values of the above hypothesis boundaries.
W_{Li} = 我; C_{Li} =Nh; $W_{R,j}$ =車禍; $C_{R,j}$ =Na;
$P_{L,}$ =S; $P_{R,}$ =NP;
Hw_L=發明; Hc_L= VC; Hw_R =車禍; Hc_R=Na;
D_L = 5; D_R = 2;

Figure 1. The syntactic structures of 5(a) and 5(b) produced by a PCFG parser.

The features used are:
$W_{L,i}$; $C_{L,i}$; $W_{R,j}$; $C_{R,j}$: The left(*i*)/right(*j*) most word and its pos category of the left/right conjunct.
$P_{L,}$; $P_{R,}$: The phrasal category of the left/right conjunct.
Hw_L ; Hc_L ; Hw_R ; Hc_R: The phrasal head and its pos category of the left/right conjunct.
D_L ; D_R: The length of the left/right conjunct.

Three types of feature patterns are used for CRF. The first type is feature patterns regarding individ-

[2] According to our experiments only 0.8% of the complex testing data and development data are failed to parse their partial structures at both sides of conjunction.

ual conjuncts. The second type is feature patterns regarding symmetry between two conjuncts. The third type is feature patterns regarding contextual properness of a conjunctive structure.

Type1: W_{Li}, W_{Li-1}, W_{Li+1}, C_{Li}, C_{Li-1}, C_{Li-2}, $C_{Li-1}C_{Li-2}$, C_{Li+1}, C_{Li+2}, $C_{Li+1}C_{Li+2}$, $C_{Li}C_{Li-1}C_{Li-2}$, $C_{Li-1}C_{Li}C_{Li+1}$, $C_{Li}C_{Li+1}C_{Li+2}$, $W_{Li}Hw_{L}$, $C_{Li}Hc_{L}$, and W_{Rj}, W_{Rj-1}, W_{Rj+1}, C_{Rj}, C_{Rj-1}, C_{Rj-2}, $C_{Rj-1}C_{Rj-2}$, C_{Rj+1}, C_{Rj+2}, $C_{Rj+1}C_{Rj+2}$, $C_{Rj}C_{Rj-1}C_{Rj-2}$, $C_{Rj-1}C_{Rj}C_{Rj+1}$, $C_{Rj}C_{Rj+1}C_{Rj+2}$, $W_{Rj}Hw_{R}$, $C_{Rj}Hc_{R}$..

Type 2: P_{L} P_{R}, $Hw_{L}Hw_{R}$, $Hc_{L}Hc_{R}$, $D_{L}D_{R}$.

Type 3: $W_{L,i+1}Hw_{Rj}$, $W_{R,j+1}Hw_{Li}$, $W_{L,1}W_{R,j}$, $W_{R,1}W_{L,j}$, $W_{L,1}W_{R,j+1}$, $W_{R,1}W_{L,j+1}$, $W_{L,1}W_{R,j}W_{R,j+1}$, $W_{R,1}W_{L,i}W_{L,i+1}$, $W_{L,1}W_{R,j-1}W_{R,j}$, $W_{R,1}W_{L,i-1}W_{L,i}$, $C_{L,i-1}Hc_{Rj}$, $C_{R,j-1}Hc_{Li}$, $C_{L,i+1}Hc_{Rj}$, $C_{R,j+1}Hc_{Li,}$ $C_{L,i}C_{L,i+1}Hc_{Rj}$, $C_{R,j}C_{R,j+1}Hc_{Li}$, $C_{L,1}C_{R,j}$, $C_{R,1}C_{L,j}$, $C_{L,1}C_{R,j+1}$, $C_{R,1}C_{L,j+1}$, $C_{L,1}C_{R,j}C_{R,j+1}$, $C_{R,1}C_{L,i}C_{L,i+1}$, $C_{L,1}C_{R,j-1}C_{R,j}$, $C_{R,1}C_{L,i-1}C_{L,i.}$

A CRF model is trained from the Sinica Treebank and estimated the probabilities of hypothesis conjunctive boundary pairs by the feature patterns listed above. The top ranked candidates are selected according to the CRF model. In general, for further improvement, a final step of semantic evaluation will be performed to select the best candidate from top-N boundary structures ranked by the CRF model, which is described in the next section.

3.2 The word-association evaluation model

For the purpose of selecting the best candidates of complex conjunctive structures, a word association evaluation model is adopted (Hsieh et al. 2007). The word-to-word association data is learned automatically by parsing texts from the Taiwan Central News Agency corpus (traditional characters), which contains 735 million characters. The syntactically dependent words-pairs are extracted from the parsed trees. The word-pairs are phrasal heads and their arguments or modifiers. Though the data is imperfect (due to some errors produced by auto-tagging system and parser), the amount of data is large enough to compensate parsing errors and reliably exhibit strength between two words/concepts.

37,489,408 sentences in CNA (Central News Agency) corpus are successfully parsed and the number of extracted word associations is 221,482,591. The word association probabilities is estimated by eq.(1).

$$P(Modify \mid Head) = \frac{freq(Head, Modify)}{freq(Head)} \tag{1}$$

"*freq*(*Head*)" means Head word frequency in the corpus and "*freq*(*Head,Modify*)" is the cooccurrence frequency of Head and Modify/Argument.

The final evaluation is done by combining three scores, i.e. (1) the probability produced by PCFG parser, (2) the scores of CRF classifier and (3) the scores of semantic evaluation. The detail is described in Section 4.2.

4 Experiments

3,484 sentences of the Sinica Treebank are used as training data. The development data and testing data are extracted from three different set of corpora the Sinica corpus, Sinorama magazines and textbooks of elementary school (Hsieh et al. 2005). They are totally 202 sentences (244 conjunctions) with 6-10 words and 107 sentences (159 conjunctions) with more than 11 words. We only test the sentences which contain the coordinate conjunction category or categories.

We adopt the standard PARSEVAL metrics (Manning et al., 1999) including bracket f-score to evaluate the performance of the tree structures of sentences and accuracies of boundary detection of conjunction structures.

4.1 Phase-1 experimental results

For the phase-1 experiments, the context-dependent rules are extracted and generalized from Sinica treebank. We then use the development data to evaluate the performances for different sets of rules selected by different threshold values. The results show that the threshold values of occurrence once and precision 70% performed best. This means any context-dependent rule with precision greater than or equal to 70% is used for the future processes. 39941 rules are in the set. In Table 1, we compare *the phase-1* result with *the baseline model* on test data. It is shown that the boundary detection precision is very high, but the recall rate is comparatively low, since *the phase-1* process cannot handle the complex cases. We also compare the processing time between *the baseline model* and *the phase-1 parsing processes* in Table 2. Marking conjunctive boundaries before parsing can limit the search range for parser and save processing time. The effect is more obvious when parsing long sentences. Because long sentences generate more ambiguous paths than shorter sentences, these surely spend much more time.

Test data	6-10 words		more than 11 words	
	Baseline	*phase1*	Baseline	*phase1*
C-boundary f-score	55.74	84.43	50.0	63.75
S-bracket f-score	72.67	84.44	71.20	79.40

Table 1. The comparison between *the baseline PCFG model* and the *phase1 parsing process* .

unit: second	6-10 words		more than 11 words	
	Baseline	*phase1*	Baseline	*phase1*
development data	14	12	34	23
test data	14	11	34	24

Table 2. The comparison of processing time between *the baseline model* and the *phase1 parsing process*.

4.2 Phase-2 experimental results

Complex cases cannot be matched by context-dependent rules at the phrase-1 which will be handled by the phase-2 algorithms mentioned in Section 3. We use the CRF++ tool (Kudo, 2006) to train our CRF model. The CRF model can produce the N-best candidates for an input conjunctive sentence. We experiment on the models of Top1-CRF and TopN-CRF where the Top1-CRF algorithm means that the final output is the best candidate produced by CRF model and the TopN-CRF means that the final output is the best candidate produced by the structure evaluation process described below.

For each N-best candidate structure, three evaluation scores is derived: (a) the probability score generated from the PCFG parser, i.e. RuleScore, (b) the probability score generated from the CRF classifier, i.e. CRF-Score, and (c) the word association score, i.e. WA-Score. We normalize each of the three scores by eq.(2):

$$normal(Score_i) = \frac{Score_i - Score_{\min}}{Score_{\max} - Score_{\min}} \qquad (2)$$

$Score_i$ means the score of the i-th candidate, and $Score_{\min}$ and $Score_{\max}$ mean the worst and the best score in the candidate set for a target conjunctive sentence. The normalized scores are between 0 and 1. After normalization, we combine the three scores with different weights:

Total Score = w1*RuleScore + w2*CRF-Score + w3*WA-Score (3)

The w1, w2 and w3 are regarded as the degree of importance of the three types of information. We use development data to determine the best combination of w1, w2, w3. Due to limit amount of development data, many local maximum and global maximum are achieved by different values of w1, w2, w3. Therefore we use a clustering algorithm to cluster the grid points of (w1, w2, w3) which produce the best performance. We then pick the largest cluster and calculate its centroid as our final weights which are shown at Table 3.

	Top N	w1	w2	w3
6-10words	N = 3	0.11	0.64	0.25
11- words	N = 3	0.18	0.76	0.06

Table 3. The best weights determined by the development data for the sentences with different lengths using the best-3 candidates.

The performance results of the testing data are shown in Table 4. In comparing with the results of the baseline model shown in Table 1, the conjunction boundary f-score increased from about 53% to 83% for the testing data. The processes also improve the overall parsing f-scores from 72% to 83%. The results of Table 4 also show that the evaluation function indeed improves the performances but marginally. However the experiments are done under the condition that the input sentences are perfectly word segmented and pos tagged. In real practices, parser may accept sentences with ambiguous word segmentation and pos tagging to avoid the error accumulation due to early commitment on word segmentation and pos tagging. Therefore parsers require much more information to resolve much more ambiguous conditions. A robust evaluation function may play a very important role. We will do more researches in the future.

		Top1CRF	TopNCRF
Development data	C-boundary f-score	85.57	89.55
	S-bracket f-score	80.10	82.34
Test data	C-boundary f-score	82.18	83.17
	S-bracket f-score	83.15	83.45

Table 4. The final results of our overall processes.

Another point worth mentioning, the performances of "CRF" (using CRF model without phase-1) and "phase1+CRF" (using CRF model after phase-1) algorithms are comparable. However "phase1+ CRF" algorithm is much more efficient, since "phase1+CRF" algorithm can determine the simple conjunctive structures by pattern matching and most of conjunctive structures are simple. On the other hand, the "CRF" model requires twice partial sentence parsing, generates candidates with CRF classifier and evaluates structure with three syntactic and semantic scores.

5 Conclusion

Conjunctive boundary detection is not a simple task. It is not only time consuming but also knowledge intensive. Therefore we propose a context-dependent rules matching approach to handle simple cases to get fast returns. For complex cases, we use a knowledge intensive divide-and-conquer approach. To resolve the problems of inadequate knowledge and data sparseness due to limit amount of structure annotated training data, we extract word/concept associations from CNA corpus.

In our experiments, the proposed model works well. Most conjunctive phrases are simple cases and can be matched by context-dependent rules and indeed avoid unnecessary calculation. Compared with the baseline method of straight forward PCFG parsing, the f-score of conjunctive boundary detection can be raised about 22%. For the complex cases, the boundaries f-score is further raised about 7% after phase-2 processes. The experimental results show that the method not only works well on boundary resolution for conjunctive phrases but also improves the total performances of syntactic parsing.

Our solutions include the rule-based method and cooperate with semantic and syntactic analyses. Therefore in the future we will try to enhance the syntactic and semantic analyses. For syntactic analysis, we still need to find more effective methods to improve the performance of our parser. For the semantic analysis, we will try to refine the word association data and discover a better semantic evaluation model.

Acknowledgements

This research was supported in part by National Digital Archives Program (NDAP, Taiwan) sponsored by the National Science Council of Taiwan under NSC Grants: NSC95-2422-H-001-031-.

References

Agarwal, Rajeev and Boggess, Lois. 1992. A Simple but Useful Approach to Conjunct Identification. In *Proceedings of 30th Annual Meeting of Association for Computational Linguistics*, pages 15-21.

Chen, Keh-Jiann, Huang, Chu-Ren, Chen, Feng-Yi, Luo, Chi-Ching, Chang, Ming-Chung, Chen, Chao-Jan and Gao, Zhao-Ming. 2003. Sinica Treebank: design criteria, representational issues and implementation. In Anne Abeille, (ed.): *Building and Using Parsed Corpora. Text, Speech and Language Technology.* 20:231-248, pages 231-248.

Hsieh,Yu-Min, Yang, Duen-Chi and Chen, Keh-Jiann. 2005. Linguistically-motivated grammar extraction, generalization and adaptation. In *Proceedings of the Second International Join Conference on Natural Language Processing (IJCNLP2005)*, pages 177-187, Jeju Island, Republic of Korea.

Hsieh, Yu-Ming, Duen-Chi Yang and Keh-Jiann Chen. 2007. Improve Parsing Performance by Self-Learning. *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 12, #2, pages 195-216.

Kurohashi, Sadao, and Nagao, Makoto. 1994. A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structure. Computational Linguistics 20(4), pages 507-534.

Kudo, Taku. 2006. (software)CRF++: Yet Another CRF toolkit *http://chasen.org/~taku/software/CRF++/.*

Lafferty, John, McCallum, Andrew, Pereira, Fernando. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning* (*ICML-01*), pages 282-289.

Lee, Yong-Hun, Kim, Mi-Young and Lee, Jong-Hyeok. 2005. Chunking Using Conditional Random Fields in Korea Texts. In *Proceedings of the Second International Join Conference on Natural Language Processing (IJCNLP2005)*, pages 155-164, Jeju Island, Republic of Korea.

Manning, Christopher D., and Schutze, Hinrich. 1999. Foundations of Statistical Natural Language processing. *The MIT Press*, Cambridge, Massachusetts.

Miller, Geroge, 1993. Introduction to WordNet: An Online Lexical Database. Princeton, CSL Report 43.

Steiner, Ilona. 2003. Parsing Syntactic Redundancies in Coordinate Structures. Poster presentation at the *European Cognitive Science Conference (EuroCogSci03).*

Van Delden, Sebastian. 2002. A Hybrid Approach to Pre-Conjunct Identification. In *Proceedings of the 2002 Language Engineering Conference* (LEC 2002), pages 72-77, University of Hyderabad, India.

Wu, Yunfang. 2003. Contextual Information of Coordinate Structure. Advances on the Research of Machine Translation, pages 103-109, Publishing house of Electronics Industry.