Automatic rule acquisition for Chinese intra-chunk relations

Qiang Zhou
Center for Speech and Language Technologies, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology
Tsinghua University, Beijing 100084, P. R. China
zq-lxd@mail.tsinghua.edu.cn

Abstract

Multiword chunking is defined as a task to automatically analyze the external function and internal structure of the multiword chunk(MWC) in a sentence. To deal with this problem, we proposed a rule acquisition algorithm to automatically learn a chunk rule base, under the support of a large scale annotated corpus and a lexical knowledge base. We also proposed an expectation precision index to objectively evaluate the descriptive capabilities of the refined rule base. Some experimental results indicate that the algorithm can acquire about 9% useful expanded rules to cover 86% annotated positive examples, and improve the expectation precision from 51% to 83%. These rules can be used to build an efficient rule-based Chinese MWC parser.

1 Introduction

In recent years, the chunking problem has become a hot topic in the communities of natural language processing. From 2000 to 2005, several different chunking-related tasks, such as text chunking (Sang and Buchholz, 2000), clause identification (Sang and Dejean, 2001), semantic role labeling (Carreras and Marquez, 2005), were defined in the CoNLL conferences. Much research has been devoted to the problem through different points of view.

Many computational linguists regard chunking as a shallow parsing technique. Due to its efficiency and robustness on non-restricted texts, it has become an interesting alternative to full parsing in many NLP applications. On the base of the chunk scheme proposed by Abney (1991) and the BIO tagging system proposed in Ramshaw and Marcus(1995), many machine learning techniques are used to deal with the problem. However, almost all the chunking systems focus on the recognition of non-overlapping cores of chunks till now, none of them care about the internal structure analysis of chunks.

In our opinion, the internal structure of a chunk, including its head and the dependency relation between head and other components, plays an important role for semantic content understanding for the chunk. They are especially useful for the languages with few morphological inflections, such as the Chinese language. Therefore, we design a multiword chunking task to recognize different multiword chunks (MWCs) with the detailed descriptions of external function and internal structure in real texts. Its main difficulty lies in the preciously identification of different lexical relationships among the MWC components. Some detailed lexical semantic knowledge is required in the task.

To deal with this problem, we proposed a rule acquisition algorithm to automatically learn a MWC rule base, under the support of a large scale annotated corpus and a lexical knowledge base. We also proposed an expectation precision index to evaluate the descriptive capabilities of the refined rule base. Some experimental results indicate that our current algorithm can acquire about 9% useful expanded rules to cover 86% annotated positive examples, and improve the expectation precision from 51% to 83%.

2 Multiword chunking task

Informally, a MWC is a chunk with two or more words, where each word links to a semantic head through different dependency relations. Four syntactic dependency relationships are used in the paper: (1) Modifier-Head relation, (2) Predicate-

Object relation,(3) Predicate-Compliment relation, (4) Coordinate relation. They can determinate the following functional position tags for each word in a MWC: (1) M--Modifier; (2) H--Head; (3) P--Predicate; (4) O--Object; (5) C--Compliment; (6) J--Coordinate constituent. Based on them, we define three topological constructions as follows:

(1) Left-Corner-Centre (LCC) construction

All the words in a chunk link to the left-corner head and form a left-head dependency structure. Its basic pattern is: H C_1 … C_n. The typical dependencies among them are Predicate-Object or Predicate-Compliment relations: $C_1 \rightarrow H$, … , $C_n \rightarrow H$. They form the following functional tag serial : P [C|O].

(2) Right-Corner-Centre (RCC) construction

All the words in a chunk link to the right-corner head and form a right-head dependency structure. Its basic pattern is: A_1 … A_n H. The typical dependencies among them are Modifier-Head relations: $A_1 \rightarrow H$, … , $A_n \rightarrow H$. They form the following functional tag serial : $\{M\}_+$ H.

(3) Chain Hooking (CH) construction

Each word in a chunk links to its right-adjacent word. All of them form a multi-head hooking chain. Its basic pattern is: H_0 H_1 … H_n, where H_i, $i \in [1,n-1]$ is the chain head in differnt levels, H_n is the semantic head of the overall chunk. The typical dependencies among them are Modifier-Head or Coordinate relations : $H_0 \rightarrow H_1$, … , $H_{n-1} \rightarrow H_n$. They form the following functional tag serial : $\{J\}_*$ or [M|J] $\{K|J\}_*$ H, where K represents the internal chain head.

We think the above three constructions can cover almost all important syntactic relations in real text sentences. Now, we can give a formal definition for a multiword chunk.

Definition: two or more words can form a multiword chunk if and only if it has one of the above three internal topological constructions.

The MWC definition builds the one-to-one corresponding between the word serials with different function tags and their dependency structure. So we can easily describe some MWCs with complex nested structures. In the paper, we add a further restriction that each MWC can only comprise the content words, such as nouns, verbs, adjectives, etc. This restriction can make us focus on the analysis of the basic content description units in a sentence.

Each MWC is assigned two tags to describe its external function and internal structure. For example, a 'np-ZX' MWC represents a noun chunk with internal modifier-head relationship. Table 1 lists all the function and relation tags used in our MWC system. The np, mp, tp, sp form as the nominal chunk set. Their typical relation tags are ZX, LN and LH. The vp and ap form as the predicate chunk set. Their typical relation tags are ZX, PO, SB and LH.

F-tags	Descriptions	R-tags	Descriptions
np	noun chunk	ZX	modifier-head relationship
vp	verb chunk	PO	verb-object relationship
ap	adjective chunk	SB	verb-compliment relationship
mp	quantity chunk	LH	Coordinate relationship
sp	space chunk	LN	chain hooking relationship
tp	time chunk		

Table 1 Function and relation tags of MWCs

The following is a MWC annotated sentence:

[tp-ZX 长期/t(long time) 以来/f(since)] ，/w 他/r(he) 为/p(for) 维护/v(safeguard) [np-ZX 世界/n (world) 和平/n(peace)] 的/u [np-ZX 崇高/a(lofty) 事业/n(undertaking)] [vp-PO 倾注/v (devote) 心血/n (painstaking)] ，/w 作出/v(make) 了/u 卓越/a(outstanding) 的/u 贡献/v (contribution) 。/w[1] (For a long time past, he has devoted all his energy into the lofty undertaking to safeguard world peace and made a outstanding contribution.) (1)

There are four MWCs in the sentence. From which, we can easily extract the positive and negative examples for a MWC rule. For example, in the sentence, we can extract a positive example: 倾注/v (devote) 心血/n (painstaking), and a negative example: 维护/v(safeguard) 世界/n (world) for the verb chunk rule : v+n→ vp-PO.

3 Automatic rule acquisition

The goal of the rule acquisition algorithm is to

[1] POS tags used in the sentence: t-time noun, f-direction, r-pronoun, p-preposition, v-verb, n-noun, u-auxilary, a-adjective, d-adverb, w-puntuation.

automatically acquire some syntactic structure rules to describe which words in which context in a sentence can be reduced to a reliable MWC, on the base of a large scale annotated corpus and a lexical knowledge base.

Each rule will have the following format: <structure description string> → <reduced tag> <confidence score>

Two types of structural rules are used in our algorithm: (1) Basic rules, where only POS tags are used in the components of a structure rule; (2) Expanded rules, where some lexical and contextual constraint is added into the structure rule string to give more detailed descriptions. The reduced tag has two kinds of MWC tags that are same as ones defined in Table 1.

Each rule consists of all the positive and negative examples covered by the rule in the annotated corpus. For the word serial matched with the structure description string of a rule, if it can be reduced as a MWC in the annotated sentence, it can be regarded as a positive example. Otherwise, it is a negative example. All of them form a special state space for each acquired rule. Therefore, the confidence score (θ) for the rule can be easily computed to evaluate the accuracy expectation to apply it in an automatic parser. Its computation formula is: $\theta = f_P / (f_P + f_N)$, where f_P is the frequency of the positive examples, and f_N is the frequency of the negative examples.

A two-step acquisition strategy is adopted in our algorithm.

The first step is rule learning. We firstly extract all basic rules with positive examples from the annotated corpus. Then, we match the extracted structure string of each basic rule in all the corpus sentences to find all possible negative examples and build state space for it. Through rule reliability computation (see the following section), we can extract all high-reliability basic rules as the final result, and all other basic rules with higher frequency for further rule refinement.

The second step is rule refining. We gradually expand each rule with suitable lexical and contextual constraint based on an outside lexical knowledge base, dynamically divide and automatically allocate its positive and negative examples into the expanded rules and form different state spaces for them. From them, we can extract all the high and middle reliability expanded rules as the final results.

At last, by combining all the extracted basic and expanded rules, we build a hierarchical acquired rule base for parser application.

Two key techniques are proposed in the algorithm:

(1) Rule reliability evaluation

The intuition assumption is that: if a rule has a higher confidence score and can cover more positive examples, then it can be regarded as a reliable rule.

Types	Decision conditions
1	● (f_P>=10) && (θ>=0.85) ● ((f_P>=5) && (f_P<10)) && (θ>=0.9) ● ((f_P>=2) && (f_P<5)) && (θ>=0.95)
2	● (f_P>=10) && (θ>=0.5) ● ((f_P>=5) && (f_P <10)) && (θ>=0.55) ● ((f_P>=2) && (f_P<5)) && (θ>=0.6) ● (f_P >0) && (θ>=0.6)
3	● (f_P >=10) && (θ>=0.1) ● ((f_P>=5) && (f_P<10)) && (θ>=0.2) ● ((f_P>=2) && (f_P<5)) && (θ>=0.3) ● (f_P>0) && (θ>=0.3)
4	All others

Table 2 Four reliability types of the acquired rules

By setting different thresholds for θ and f_P, we can classify all acquired rules into the following four types of rule sets: (1) high-reliability (HR) rules; (2) middle-reliability (MR) rules; (3) low-reliability rules; (4) Useless and noise rules. Table 2 shows different decision conditions for them in our current algorithm. Based on this uniform evaluation standard, we can easily extract effective rules from different acquired rule base and quickly exclude useless noise rules.

(2) Rule expansion and refinement

When a rule is not reliable enough, the expansion step is set off: new knowledge is added to the rule in order to constrain it. The purpose is to dynamically divide the state space of the rule and reduce the proportion of negative examples covered by the current rule. For every annotated positive or negative example, our expansion strategy is as follows:

Firstly, we expand a rule description through looking up different lexical knowledge base. For the verb chunks with LCC constructions, we use the following lexical constraint: (1) Lexical-syntactic relation pairs, (2) Subcategory frame of

head verb. For the noun chunks with RCC and CH constructions, we use the following lexical constraint: (1) Lexical-syntactic relation pairs, (2) Semantic class of head noun.

Secondly, we expand a rule description example with or without lexical constraint through looking up its left and right adjacent contexts. For each rule waiting for expansion, we add its left-adjacent POS tag, right-adjacent POS tag, left and right adjacent POS tag to form three expanded rule with contextual constraint.

For example, for the positive example "倾注/v (devote) 心血/n (painstaking) " of "v+n" rule in the above sentence (1), we can get the following expanded rules:

- v(WC-L)+n(WC-R) // + v-n relationship pair
- v(winl:VNPLIST)+n // + verb subcate frame
- n__v+n // + left POS constraint
- v+n__w // +right POS constraint
- n__v+n__w // +l and +r POS constraint

They can be put into the state space pool as the expanded rules with positive example information for frequency calculation.

Unlike the *information-gain* measure used in FOIL system (Quinlan, 1990), we do not impose any criteria for selecting different knowledge. All the suitable expanded rules are selected through the final confidence score evaluation indexes.

4 Experimental results

All the news files with about 200,000 words in the Chinese treebank TCT (Zhou, 2004) were selected as the experimental data. They were separated into two data sets: (1) training set, which consists of about 80% data and is used for rule acquisition; (2) test set, which consists of about 20% data and is used for parser evaluation.

Then we automatically extracted all the MWCs from the annotated trees and built two MWC banks. Among them, 76% are noun chunks and verb chunks. They are the key points for rule acquisition and parsing application. In the training set, about 94% verb chunks are two-word chunks. But for noun chunks, the percentage of two-word chunks is only 76%. More than 24% noun chunks comprise three or more words. The complexities of noun chunks bring more difficulties for rule acquisition and automatic MWC parsing.

We also used the following lexical knowledge base for rule expansion and refinement: (1) Lexical relationship base. It consists of 966953 lexical pairs with different syntactic relationships. All the data are extracted from 4 different language resources. (2) Verb subcategory data. It consists of 5712 verbs with the "v+np" subcat frames and 1065 verbs with the "v+vp" subcat frames. All the data are extracted from a Chinese grammatical dictionary (Yu and al., 1998). (3) Noun thesaura data. It consists of 26906 nouns annotated with the different semantic types All the data are extracted from Hownet-2000[2].

4.1 Rule base acquisition

We ran our algorithm on the above language resources and obtained the following results.

In the rule learning stage, we extracted 735 basic rules from the training set. After reliability evaluation, we obtained 61 HR rules and 150 less reliable rules for further refinement. Although these 211 rules only make up 29% of all the 735 acquired rules, they cover about 97% positive examples in the training set. Thus, almost all the useful information can be reserved for further rule expansion and refinement.

In the rule refining stage, 47858 rules were expanded from the 150 basic rules. Among them, all 2036 HR and 2362 MR rules were selected as the final results. They make up about 9% of all the expanded rules, but cover 86% positive examples. It indicates the effectiveness of our current rule acquisition algorithm.

In order to evaluate the descriptive capability of the acquired rules objectively, we proposed an expectation precision (*EP*) index to estimate the parsing accuracy when we apply the acquired rules to all the positive examples in the training set. Its computation formula is as follows:

$$EP = \sum_{i=1}^{N} (f_{Pi} * \theta_i) / \sum_{i=1}^{N} f_{Pi}$$

where N is the total number of the rules in a rule base, f_{Pi} and θ_i are the positive example frequency and confidence score of the i^{th} rule in the rule base. An intuition assumption behind the *EP* definition is that a rule base with higher *EP* index will imply its better descriptive capability for some special linguistic phenomena. Therefore, its better parsing performance in a rule-based parser can be expected. To prove this assumption, we designed a

[2] The data is available in http://www.keenage.com

simple comparison experiment to analyze the improvement effects of different lexical and contextual constraint used in our expanded rules.

We divided all 150 basic rules into 4 subsets, according to their different internal structure characteristics: (1) Noun chunks with RCC and CH constructions; (2) Verb chunks with LCC constructions; (3) Verb chunks with RCC constructions; (4) All other MWCs.

The rules in the subset 1 and 2 cover majority of the positive examples in the training set. They have complex internal structures and lexical relations. So we applied the lexical knowledge base and contextual constraint to expand them. Comparatively, the rules in subset 3 and 4 have simpler structures, so we only used the contextual constraint to expand them.

Table 3 shows the *EP* indexes of these rule subsets before and after rule refining. For all 150 basic rules, after rule expansion and refinement, the *EP* index was improved about 65%. For the simpler structure rules in subset 3 and 4, just the application of contextual constraint can bring dramatic improvement in the *EP* index. It indicates the importance of the local contextual information for multiword chunk recognition.

Sub set	Rule sum	Covered positive examples	EP before expansion (%)	EP after expansion (%)
1	51	13689	52.70	81.40
2	20	8859	45.14	80.56
3	24	2342	28.12	93.27
4	55	3566	66.85	93.22
Total	150	28456	50.56	83.36

Table 3 Descriptive capability analysis of different kinds of expanded rule sets

For the major subset 1 and 2, *EP* index also shows great improvement. It increased about 54% and 78% in the subset 1 and 2 respectively. As we can see, the applying effects of lexical and contextual constraint on the verb chunks were superior to that on the noun chunks. Two factors contribute to this phenomenon. First, the simpler internal structures of most verb chunks guarantee the availability of almost all corresponding lexical relationship pairs. Second, most lexical pairs used in verb chunks have stronger semantic relatedness than that in noun chunks.

4.2 Parsing performance evaluation

Based on the rule base automatically acquired through the above algorithm, we developed a rule-based MWC parser to automatically recognize different kinds of MWCs in the Chinese sentences after word segmentation and POS tagging. Through θ-based disambiguation technique, the parser can output most reliable MWCs in the disambiguated region of a sentence and keep some ambigous regions with less reliable MWC structures to provide multiple selection possibilities for a full syntactic parser. Some detailed information of the parser can be found in (Zhou, 2007).

We used three commonly-used indexes : precision, recall and F-measure to evaluate the performance of the parser. Two different criteria were set to determinate the correctness of a recognized MWC. (1) 'B+F+R' criterion : It must have the same left and right boundaries, function tag and relation tag as that of the gold standard. (2) 'B+F' criterion : It must have the same left and right boundaries, function tag as that of the gold standard.

Table 4 shows the experimental results under the disambigutated regions, which cover 95% of the test data.

Type	'B+F+R' criterion	'B+F' criterion
np	75.25/75.76/75.50	83.68/84.25/83.97
vp	83.23/81.46/82.34	87.35/85.49/86.41
mp	94.89/95.26/95.08	94.89/95.26/95.08
ap	93.99/97.33/95.63	93.99/97.33/95.63
tp	92.75/88.18/90.40	93.52/88.92/91.16
sp	78.76/86.41/82.41	79.65/87.38/83.33
Total	81.76/81.44/81.60	87.01/86.67/86.84

Table 4 Open test results (P/R/F-m, %) under the disambiguated regions

The differences of F-measures among three MWC subsets, i.e. noun chunks, verb chunks and other chunks, show interesting positive association with the differences of their *EP* indexes listed in the previous sections. When we apply the acquired rule base with higher *EP* index in the rule-based parser, we can get better parsing performance. It indicates that *EP* value can be used as an important objective index to evaluate the descriptive capability of the rule base automatically acquired for large scale annotated corpus.

The lower F-measure of noun and verb chunk

under 'B+F+R' criterion shows the difficulty for lexical relation recognition, especially for the complex noun chunks. There are still much improvement room in future research.

5 Related work

In the area of chunking rule acquisition and refinement, several approaches have been proposed. Cardie and Pierce(1999) explored the role of lexicalization and pruning of grammars for base noun phrase identification. Their conclusion is that error-driven pruning is a remarkably robust method for improving the performance of grammar rules. Dejean(2002) proposed a top-down inductive system, ALLis, for learning and refining linguistic structures on the base of contextual and lexicalization knowledge extracted from an annotated corpus. Choi et all(2005) proposed a method for automatically extracting partial parsing rules from a tree-annotated corpus using decision tree induction. The acquired grammar is similar to a phrase structure grammar, with contextual and lexical information, but it allows building structures of depth one or more.

All these researches prove the important role of lexical and contextual information for improving the rule descriptive capability. However, the lexical information used in these systems is still restricted in the lexical head of a constituent. None of the lexical relationship knowledge extracted from the annotated corpus or other outside language resources has been applied. Therefore, the room for improvement of the rule descriptive capability is restricted to a certain extent.

6 Conclusions

Three main contributions of the paper are summarized as follows. (1) We design a new multiword chunking task. Based on the topological structure definition, we establish the built-in relations between multiword chunk examples in annotated corpus and lexical relationship pairs in outside lexical knowledge base. (2) We propose an efficient algorithm to automatically acquire hierarchical structure rules from large-scale annotated corpus. By introducing different kinds of lexical knowledge coming from several different language resources, we set up an open learning environment for rule expansion and refinement. (3) We propose an expectation precision index to evaluate the descriptive capability of the refined rule base. Experimental results show that it has stronger positive association with the F-measure of parser performance evaluation.

Acknowledgements. The research was supported by NSFC (Grant No. 60573185, 60520130299). Thank the comments and advice of the anonymous reviewers.

References

Steven Abney. 1991. Parsing by Chunks. In *R. Berwick, S. Abney and C. Tenny (eds.) Principle-Based Parsing, Kluwer Academic Publishers.*

Claire Cardie and D. Pierce. 1999. The Role of Lexicalization and Pruning for Base Noun Phrase Grammars. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99).

X. Carreras and L. M`arquez. 2005. Introduction to the conll-2004 shared tasks: Semantic role labeling. In Proc. of CoNLL-2005.

Myung-Seok Choi, Chul Su Lim, and Key-Sun Choi. 2005. Automatic Partial Parsing Rule Acquisition Using Decision Tree Induction. In *R. Dale et al. (Eds.). Proc. of IJCNLP 2005, Seoul, Korea* . p143–154.

Herve Dejean. 2002 Learning rules and their exceptions. *Journal of Machine Learning Research*, 2002: 669–693.

J R. Quinlan 1990. Learning logical definitions from relations. *Machine Learning*, 5:239–266.

L Ramshaw and M Marcus. 1995.Text chunking using transformation-based learning. In *Proc. of the Third Workshop on Very Large Corpora*, p82-94.

Erik F. Tjong Kim Sang and S. Buchholz 2000 Introduction to CoNLL-200 Shared Task: Chunking. In *Proc. of CoNLL-2000 and LLL-2000*. Lisbon. p127-132.

Erik F. Tjong Kim Sang and H. Déjean 2001. Introduction to the CoNLL-2001 Shared Task: Clause Identification. In *Proc. of CoNLL-2001*, Toulouse, France, p53-57.

Shiwen Yu, Xuefeng Zhu, et al. 1998 *A Complete Specification of the Grammatical Knowledge-base of Contemporary Chinese*. Tsinghua University Press. (in Chinese)

Qiang Zhou 2004. Annotation scheme for Chinese Treebank. *Journal of Chinese Information Processing*, 18(4): 1-8. (in Chinese)

Qiang Zhou. 2007. A rule-based Chinese chunk parser. In *Proc. Of ICCC-2007,* furthercoming.