The Effect of Rhythm on Structural Disambiguation in Chinese

Honglin Sun **Dan Jurafsky**
Center for Spoken Language Research
University of Colorado at Boulder
{honglin.sun, jurafsky}@colorado.edu

Abstract

The length of a constituent (number of syllables in a word or number of words in a phrase), or *rhythm,* plays an important role in Chinese syntax. This paper systematically surveys the distribution of rhythm in constructions in Chinese from the statistical data acquired from a shallow tree bank. Based on our survey, we then used the rhythm feature in a practical shallow parsing task by using rhythm as a statistical feature to augment a PCFG model. Our results show that using the probabilistic rhythm feature significantly improves the performance of our shallow parser.

1 Introduction

Syntactic research indicates that prosodic features, including stress, rhythm, intonation, and others, have an impact on syntactic structure. For example, normally in a coordination construction like "A and B", A and B are interchangeable, that is to say, you can say "B and A" and the change of word order does not change the meaning. However, sometimes A and B are not interchangeable. Quirk et al.(1985) gives the following examples:

man and woman * woman and man
ladies and gentleman *gentleman and ladies

Obviously, the examples above cannot be explained by gender preference. A reasonable explanation is that the length of the words (perhaps in syllables) is playing a role; the first constituent tends to be shorter than the second constituent.

This feature of the length in syllables of a constituent plays an even more important role in Chinese syntax than in English (Feng, 2000). For example, in the verb-object construction in Chinese, there is a preference for the object to be equal to or longer than the verb. Thus while both "种"(plant) and "种植"(plant) are verbs and have the same meaning, "种/plant 树/tree" is grammatical while "种 植/plant 树/tree" is ungrammatical. However, both verbs allow bi-syllabic nouns as objects (e.g., "果树"(fruit tree), "棉花"(cotton) etc.). The noun phrases formed by "noun + verb" give us another example in which rhythm feature places constraints on syntax, as indicated in the following examples (ungrammatical with *):

棉花/cotton 种植/planting
*棉花/cotton 种/planting
*花/flower 种植/planting
*花/flower 种/planting

“棉花/cotton 种植/planting” is grammatical but “棉花/cotton 种/planting” , “花/flower 种植/planting” and “花/flower 种/planting” are all ungrammatical, although “棉花/cotton” and “花/flour” , “种植/planting” and “种/planting” have the same POS and the same or similar meaning. The only difference lies in that they have different number of syllables or different length.

This paper systematically surveys the effect of rhythm on Chinese syntax from the statistical data from a shallow tree bank. Based on the observation that rhythm places constraints on syntax in Chinese, we try to deploy a feature based on rhythm to improve disambiguation in a probabilistic parser by mixing the rhythm feature into a statistical parsing model.

The rest of the paper is organized as follows: we present specific statistical analyses of rhythm feature in Chinese syntax in Section 2. Section 3 introduces the content chunk parsing which is the task in our experiment. Section 4 presents the statistical model used in our experiment in which a probabilistic rhythm feature is integrated. Section 5 gives the experimental results and finally Section 6 draws some conclusions.

2 Analysis of Rhythmic Constraints

We divide our analysis of the use of rhythm in Chinese phrases into two categories, based on two types of phrases in Chinese: (1) simple phrases, containing only words, i.e. all the child nodes are POS tag in the derivation tree; and (2) complex phrases in which at least one constituent is a phrase itself, i.e. it has at least one child node with phrase type symbol (like NP, VP) in its derivation tree.

Below we will give the statistical analysis of the distribution of rhythm feature in different constructions from both simple and complex phrases. The corpus from which the statistical data is drawn contains 200K words of newspaper text from the People’s Daily. The texts are word-segmented, POS tagged and labeled with content chunks. The content chunk is a phrase containing only content words, akin to a generalization of a BaseNP. These content chunks are parsed into binary shallow trees. More details about content chunks can be found in Section 3.

2.1 Rhythm feature in simple phrases

Simple phrases contain two lexical words (since, as discussed above, our parse trees are binary). The rhythm feature of each word is defined to be the number of syllables in it. Thus the rhythm feature for a word can take on one of the following three values: (1) monosyllabic; (2) bi-syllabic; and (3) multi-syllabic, meaning with three syllables or more.

Since each binary phrase contains two words, the set of rhythm features for a simple phrase is:

F = { (0，0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2) }

where 0, 1, 2 represent monosyllabic, bi-syllabic and multi-syllabic respectively.

In the following sections, we will present three case studies on the distributions of rhythm feature in different constructions: (1) verbs as modifier or head in NP; (2) the contrast between NPs and VPs formed by “ verb + noun” sequences; (3) “noun + verb” sequences.

2.1.1 Case 1: Verb as modifier/head in NP

In Chinese, verbs can function as modifier or head in a noun phrase without any change of forms. For example, in “果树/fruit tree 栽培/growing”, “栽培” is the head while in “栽培/growing 技术/technique”, “栽培” is a modifier. However, in such constructions, there are strong constraints on the length of both verbs and nouns. Table 1 gives the distributions of the rhythm feature in the rule “NP -> N V”(‘N’ and ‘V’ represent noun and verb respectively) in which the verb is the head and “NP -> V N” in which the verb is a modifier.

Table 1 Distribution of rhythm feature in NP with verb as modifier or head

	[0,0]	[0,1]	[0,2]	[1,0]	[1,1]	[1,2]	[2,0]	[2,1]	[2,2]	Total
NP -> V N	0	4	0	0	1275	4	0	88	0	1371
NP -> N V	13	10	0	401	2328	91	0	44	2	2889

Table 2 Distribution of rhythm feature in NP and VP formed by "V N"

	[0,0]	[0,1]	[0,2]	[1,0]	[1,1]	[1,2]	[2,0]	[2,1]	[2,2]	Total
VP -> V N	826	640	49	80	1221	121	0	11	1	2777
NP -> V N	13	10	0	401	2328	91	0	44	2	2889

Table 3 Distribution of rhythm feature in phrases formed by "N V" sequence

	[0,0]	[0,1]	[0,2]	[1,0]	[1,1]	[1,2]	[2,0]	[2,1]	[2,2]	Total
NP -> N V	0	4	0	0	1275	4	0	88	0	1371
NC -> N V	384	578	42	1131	3718	143	90	435	15	6536
S -> N V	28	1	2	17	347	22	2	43	8	470

Table 1 indicates that in both rules, the rhythm pattern [1,1], ie. "bi-syllabic + bi-syllabic", prevails. In the rule "NP -> V N", this pattern accounts for 93% among the nine possible patterns while in the rule "NP -> N V", this pattern accounts for 81%. We can also find that in both cases, [0,2] and [2,0] are prohibited, that is to say, both verbs and nouns cannot be longer than two syllables.

2.1.2 Case 2: Contrast between NP and VP formed by "V N" sequence

The sequence "V N"("verb + noun") can constitute an NP or a VP. The rhythm patterns in the two types of phrases are significantly different, however, as shown in Table 2. We see that in the NP case, verbs are mainly bi-syllabic. The total number of examples with bi-syllabic verbs in NP is 2820, accounting for 98% of all the cases. On the other hand, mono-syllabic verbs are less likely to appear in this position. The total number of examples with mono-syllabic verbs in NP is 23, accounting for only 0.8% of all the cases. That is to say, the likelihood of bi-syllabic verbs appearing in this syntactic position is 122 times the likelihood of mono-syllabic verbs. On the other hand, there is no big difference between bi-syllabic verbs and mono-syllabic verbs in the VP formed by "V + N". The ratios of bi-syllabic and mono-syllabic verbs in VP are 48 % and 55% respectively. The statistical facts tell us that for a "verb + noun" sequence, if the verb is not bi-syllabic then it is very unlikely to be an NP. Figure 1 depicts more clearly the difference between NP and VP formed by "V N" sequence in the distribution of rhythm feature.

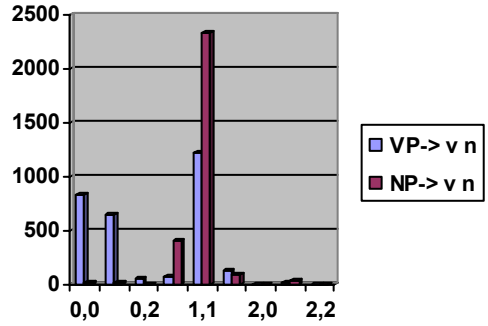


Figure 1 Distributions of rhythm feature in NP and VP formed by "verb + noun"

.

2.1 3 Case 3: "N V" sequence

An "N V"("noun + verb") sequence can be mainly divided into three types by the dominating phrasal category:

(1) NP(noun phrase), e.g. "果树/fruit tree 栽培/growth";

(2) S(subject-verb construction), e.g. " 彩旗/colored flag 飘扬/flutter";

(3)NC(non-constituent), eg. “经济/economy 发展/develop” in “中国/China 的/DE 经济/economy 发展/develop 得/DE 很/very 快/fast”. (‘China’s economy develops very fast’)

Table 3 gives the distribution of rhythm feature in the three types of cases.

We see in Table 3, in rule “NP -> N V”, that the verb cannot be mono-syllabic since the first row is 0 in all the patterns in which verb is mono-syllabic([0,0], [1,0],[2,0]). The “bi-syllabic + bi-syllabic” ([1,1]) pattern accounts for 93% (1275/1371) of the total number. Let’s look at the cases with mono-syllabic verbs in all the three types. The total number of such examples is 1652 in the corpus (adding all the numbers in columns [0,0], [1,0] and [2,0] on the three rows). Among these 1652 cases, there is not one example in which the “N V” is an NP. The sequence has a probability of 3%(47/1652) to be an S and 97 %(1605/1652) of being an NC(non-constituent).

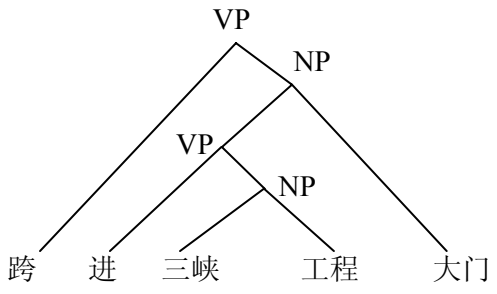
2.2 Rhythm feature in complex phrases

Just as we saw with two word simple phrases, the rhythm feature also has an effect on complex phrases where at least one component is a phrase, i.e. spanning over two words or more. For example, for the following fragment of a sentence:

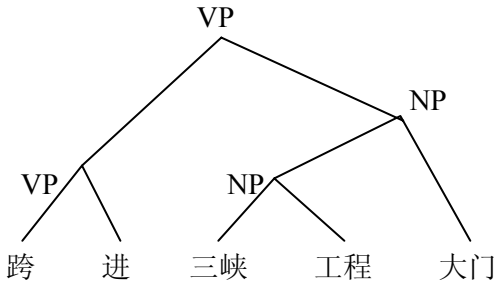
跨/stride 进/into 三峡/the Three Gorges 工程/project 大门/gate

‘enter into the gate of the Three Gorges Project’

according to PCFG, the parse as indicated in Figure 2 (a) is incorrectly assigned the greatest



(a)



(b)

Figure 2 (a) incorrect parse and (b) correct parse

probability but the correct parse is that given in Figure 2 (b). One major error in (a) is that it applies the rule “NP-> VP N” (i.e. “进 三峡 工程” modifying “大门”). This rule has 216 occurrences in the corpus, of which 168 times it contains a VP of 2 words, 30 times a VP of 3 words and 18 times a VP of more than 3 words. These statistics indicate that this rule prefers to choose a short VP acting as the modifier of a noun, as in “NP(VP(种/grow 粮/grain) 大户/large family)” and “NP(VP(学/learn 雷锋/Lei Feng) 标兵/model)”. But in the example in Figure 2(a), the VP contains 3 words, so it is less likely to be a modifier in an NP.

When a phrase works as a constituent in a larger phrase, its rhythm feature is defined as the number of words in it. Thus a phrase may take on one of the three values for the rhythm feature: (1) two words; (3) three words; and (3) more than three words. Similar to that in the simple phrases, we may use 0, 1, 2 to represent the three values respectively. Therefore, for every construction containing two constituents, its rhythm feature can be described by a 3×3 matrix uniformly. For example, in the examples for rule “NP -> VP N” above, the feature value for “NP(VP(种/grow 粮/grain) 大户/large family)” is [0, 1] in which 0 indicates the VP contains 2 words and 1 represents that the noun is bi-syllabic. The rule helps to interpret the meaning of the feature value, i.e. the

value is for a word or a phrase. For example, for rule "VP -> V N", feature value [0, 1] means that the verb is mono-syllabic and the noun is bi-syllabic, while for rule "NP-> VP N", feature [0,1] means that the VP contains two words and the noun is bi-syllabic.

3 Content Chunk Parsing

We have chosen the task of content chunk parsing to test the usefulness of our rhythm feature to Chinese text. In this section we address two questions: (1) What is a content chunk? (2) Why are we interested in content chunk parsing?

A content chunk is a phrase formed by a sequence of content words, including nouns, verbs, adjectives and content adverbs. There are three kinds of cases for the mapping between content word sequences and content chunks:

(1) A content word sequence is a content chunk. A special case of this is that a whole sentence is a content chunk when all the words in it are content words, eg. [[前景/Prospect 公司/company]NP [推出/release [高级/advanced [电脑/computer [排版/typesetting 系统/system]NP]NP]NP]VP ('Prospect Company released an advanced computer typesetting system.').

(2) A content word sequence is not a content chunk. For example, in "中国/China 的/AUX 经济/economy 发展/develop 得/AUX 很/very 快/fast"('China's economy develops very fast.'), "经济/economy 发展/develop" is a content word sequence, but it's not a phrase in the sentence.

(3) A part of a content word sequence is a content chunk. For example, in "私营/private 经济/economy 发展/develop 的/AUX 势头/trend 很/very 好/good"('The developmental trend of private economy is very good.'), "私营/private 经济/economy 发展/develop" is a content word sequence, but it's not a phrase; only "私营/private 经济/economy" in it is a phrase.

The purpose of content chunk parsing is to recognize phrases in a sequence of content words. Specifically speaking, the content chunking contains two subtasks: (1) to recognize the maximum phrase in a sequence of content words; (2) to analyze the hierarchical structure within the phrase down to words. Like baseNP chunking(Church, 1988; Ramshaw & Marcus 1995), content chunk parsing is also a kind of shallow parsing. Content chunk parsing is deeper than baseNP chunking in two aspects: (1) a content chunk may contain verb phrases and other phrases even a full sentence as long as the all the components are content words; (2) it may contain recursive NPs. Thus the content chunk can supply more structural information than a baseNP.

The motives for content chunk parsing are two-fold: (1) Like other shallow parsing tasks, it can simplify the parsing task. This can be explained in two aspects. First, it can avoid the ambiguities brought up by functional words. In Chinese, the most salient syntactic ambiguities are prepositional phrases and the "DE" construction. For prepositional phrases, the difficulty lies in how to determine the right boundary, because almost any constituent can be the object of a preposition. For "DE" constructions, the problem is how to determine its left boundary, since almost any constituent can be followed by "DE" to form a "DE" construction. Second, content chunk parsing can simplify the structure of a sentence. When a content chunk is acquired, it can be replaced by its head word, thus reducing the length of the original sentence. If we get a parse from the reduced sentence with a full parser, then we can get a parse for the original sentence by replacing the head-word nodes with the content chunks from which the head-words are extracted. (2) The content chunk parsing may be useful for applications like information extraction and question answering. When using template matching, a content chunk

may be just the correct level of shallow structure for matching with an element in a template.

4 PCFG + PF Model

In the experiment we propose a statistical model integrating probabilistic context-free grammar (PCFG) model with a simple probabilistic features (PF) model. In this section we first give the definition for the statistical model and then we will give the method for parameter estimation.

4.1 Definition

According to PCFG, each rule r used to expand a node n in a parse is assigned a probability, i.e.:

$$P(r(n)) = P(\beta \mid A) \quad (1)$$

where A -> β is a CFG rule. The probability of a parse T is the product of each rule used to expand each node n in T:

$$P(T \mid S) = \prod_{n \in T} P(r(n)) \quad (2)$$

We expand PCFG by the way that when a left hand side category A is expanded into a string β, a feature set FS related to β is also generated. Thus, a probability is assigned for expansion of each node n when a rule r is applied:

$$P(r(n)) = P(FS, \beta \mid A) \quad (3)$$

where A -> β is a CFG rule and FS is a feature set related to β. From Equation (3) we get:

$$P(r(n)) = P(FS \mid \beta, A) * P(\beta \mid A) \quad (4)$$

where P(FS| β, A) is probabilistic feature(PF) model and P(β | A) is PCFG model. PF model describes the probability of each feature in feature set FS taking on specific values when a CFG rule A -> β is given. To make the model more practical in parameter estimation, we assume the features in feature set FS are independent from each other, thus:

$$P(FS \mid \beta, A) = \prod_{Fi \in FS} P(Fi \mid \beta, A) \quad (5)$$

Under this PCFG+PF model, the goal of a parser is to choose a parse that maximizes the following score:

$$Score(T \mid S) = \arg\max_{T} \prod_{i=1}^{n} P(FS_i, \beta_i \mid A_i) \quad (6)$$

Our model is thus a simplification of more sophisticated models which integrate PCFGs with features, such as those in Magerman(1995), Collins(1997) and Goodman(1997). Compared with these models, our model is more practical when only small training data is available, since we assume the independence between features. For example, in Goodman's probabilistic feature grammar (PFG), each symbol in a PCFG is replaced by a set of features, so it can describe specific constraints on the rule. In the PFG model the generation of each feature is dependent on all the previously generated features, thus likely leading to severe sparse data problem in parameter estimation. Our simplified model assumes independence between the features, thus data sparseness problem can be significantly alleviated.

4.2 Parameter Estimation

Let F be a feature associated with a string β, where the possible values for F are $f_1, f_2, \ldots, f_n$, E is the set of observations of rule A → β in the training corpus, and thus E can be divided into n disjoint subsets: $E_1, E_2, \ldots, E_n$, corresponding to $f_1, f_2, \ldots, f_n$ respectively. The probability of F taking on a value of f_i given A → β can be estimated as follows, according to MLE:

$$P(F = f_i \mid \beta, A) = \frac{|E_i|}{|E|} \quad (7)$$

This indicates that feature F adds constraints on CFG rule A → β by dividing Ω, the state space of A → β, into n disjoint subspaces $\omega_1, \omega_2, \ldots, \omega_n$, and each case of F taking a value of f_i given A → β is viewed as a random event.

5 Experimental Results

5.1 Training and Test Data

A Chinese corpus of 200K words extracted from the People's Daily are segmented, POS-tagged and hand-labeled with content chunks in which all the trees are binary. The corpus is divided into two parts: (1) 180K for training set and (2) 20K for test set.

5.2 Metrics and results

We take two kinds of criteria to measure the system's performance: labeled and unlabeled. According to the labeled criterion, a recognized phrase is correct only if a phrase with the same starting position, ending position and the same label is found in the gold standard. According to the unlabeled criterion, a recognized phrase is correct as long as a phrase with the same starting position and ending position is found in the gold standard.

Table 4 Experimental Results

	Labeled			Unlabeled		
	P	R	F	P	R	F
PCFG	49.91	64.96	56.45	53.33	80.73	65.66
PCFG+RF in simple phrases	53.25	68.46	59.90	57.46	81.21	67.30
PCFG +RF in all the phrases	56.47	72.08	63.33	60.07	83.57	69.90

Table 5 Effect of rhythm feature on structural disambiguation

Word sequence	Rule	P(β|A)	RF	P(RF=[0,1]|A,β)	P(RF=(0,1),β)|A)
国 捐躯 country sacrifice	NC→ N V	0.120273	[0,1]	0.08843	0.010636
国 捐躯	S → N V	0.161679	[0,1]	0.00292	0.000344
国 捐躯	NP→ N V	0.063159	[0,1]	0.00213	0.000184
国 捐躯	V → N V	0.011573	[0,1]	0.0	0.0

Within each criterion, precision, recall and F-measure are given as metrics for the system's performance. Precision represents how many phrases are correct among the phrases recognized, recall represents how many phrases in the gold standard are correctly recognized, and F-measure is defined as follows:

$$F-measure=\frac{\mathrm{Pr}ecision\times \mathrm{Re}call\times 2}{\mathrm{Pr}ecision+\mathrm{Re}call}$$

Table 4 gives the experimental results in three different conditions: the first row gives the result of PCFG model; the second row gives the result of PCFG model integrated with rhythm feature model (RF) where only the features of simple phrases are considered; the last row gives the result of PCFG model plus RF where the rhythm features in all the phrases are considered. The results indicate that the rhythm features in both simple and complex phrases contribute to the improvement of performance over PCFG model. We see that the rhythm feature improves the labeled F-measure 6.88 percent and the unlabeled F-measure 4.24 percent over the unaugmented PCFG model.

5.3 Effect of rhythm feature on parsing

The experiment shows that the rhythm feature can help the performance of a parser in Chinese. Specifically, the effects of rhythm feature on parsing are shown in two ways:

(1) Help for the disambiguation of phrasal type.

Table 5 shows the difference of the results between PCFG model and PCFG + RF model for the sequence "国/country 捐躯/sacrifice" in the sentence "该/the 校/school 有/have 900 学子

/students 为/for 国/country 捐躯/sacrifice“ (`900 students from this school gave their lives for their country’).

In the sentence above, “国/country” is the object of preposition “为/for”, “国/country 捐躯/sacrifice” is not a constituent. But the unaugmented PCFG model incorrectly parses it as a S(subject-predicate construction). Contrarily, according to PCFG+RF model, the type with greatest probability is the (correct) NC(non-constituent) parse.

(2) Help for pruning.

Let’s give an example to explain it. For the sentence “解决/solve 居民/resident 吃/eat 菜/vegetable 问题/problem 十分/very 困难/difficult”(‘It’s very difficult to solve the vegetable problem for the residents.’), the number of edges generated by the PCFG is 1236, but the number decreases to 348 after the rhythm feature is applied, thus pruning 73% of the edges. As indicated in Table 1, in the rule “NP -> N V”, P(RF = [1,0]) = 0, so “[居民/N 吃/V]NP” is pruned after adding RF. Similarly, in rule “NP -> V N”, P(RF = [0, 1]) = 0.003, so “[吃/V 菜/N]NP” is pruned since it has very low probability. With these two edges pruned, more potential edges containing them will not be generated.

6 Conclusion

In this paper, we systematically survey the distribution of rhythm (number of syllables per word or numbers of words per phrase for a constituent) in different constructions in Chinese. Our analysis suggests that rhythm places strong constraints on Chinese syntax. Based on this observation, we used the rhythm feature in a practical shallow parsing task in which a PCFG model is augmented with a probabilistic representation of the rhythm feature. The experimental results show that the probabilistic rhythm feature aids in disambiguation in Chinese and thus helps to improve the performance of a Chinese parser. We can expect that the performance of the parser may further improve when more features are considered under the probabilistic feature (PF) model.

Acknowledgments

This research was partially supported by the NSF, via a KDD extension to NSF IIS-9978025.

References

Church,,K.,1988.A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp.136-143.

Collins, M. 1997. Three generative lexicalized models for statistical parsing, in *Proceedings of the 35th Annual Meeting of the ACL*, pp. 16-23.

Feng, Shengli. 2000. *The Rhythmic syntax of Chinese*(in Chinese), Shanghai Education Press.

Goodman, J. 1997. Probabilistic Feature Grammars, In *Proceedings of the International Workshop on Parsing Technologies*, September 1997

Magerman, D. 1995. Statistical decision-tree models for parsing, in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp.276-283.

Quirk et al. 1985. *A Comprehensive Grammar of English Languge*, Longman.

Ramshaw L., and Marcus M. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*.pp.86-95.