A Quasi-Dependency Model for Structural Analysis of Chinese BaseNPs*

Zhao Jun Huang Changning
Department of Computer Science & Technology,
The State Key Lab of Intelligent Technology & Systems,
Tsinghua University, Beijing, China, 100084
Email: zj@s1000e.cs.tsinghua.edu.cn, Hcn@tsinghua.edu.cn

Abstract: The paper puts forward a quasi-dependency model for structural analysis of Chinese baseNPs and a MDL-based algorithm for quasi-dependency-strength acquisition. The experiments show that the proposed model is more suitable for Chinese baseNP analysis and the proposed MDL-based algorithm is superior to the traditional ML-based algorithm. The paper also discusses the problem of incorporating the linguistic knowledge into the above statistical model.

1. Introduction

The concept of baseNP is initially put forward by Church. In English, baseNP is defined as 'simple non-recursive noun phrases', which means that there is no sub-noun-phrases contained in a baseNP[1]. But the definition can not meet the needs in Chinese information retrieval. The noun phrases such as "自然(natural) 语言(language) 处理(process)", "亚洲(Asian) 金融(finance) 危机(crisis)" and "政治(political) 体制(system) 改革(reformation) 进程(process)" are critical for information retrieval, but they are not non-recursive noun phrases.

In Chinese, the attribute of noun phrases can be classified into three types, that is restrictive attributes, distinctive attributes and descriptive attributes, among which the restrictive attributes have agglutinative relation with the heads. The paper defines the Chinese baseNP using the restrictive attributes.

【 Definition 1 】 Chinese baseNP (hereafter abbreviated as baseNP):

baseNP → baseNP + baseNP
baseNP → baseNP + N | VN
baseNP → restrictive-attribute + baseNP
baseNP → restrictive-attribute + N | VN
restrictive-attribute → A | B | V | N | S | X | (M+Q)

Where, the terminal symbols A, B, V, N, VN, S, X, M, Q stand for respectively adjective, distinctives, verbs, nouns, norminalized verbs, locatives, non-Chinese string, numerals and quantifiers.

According to the definition, noun phrases falls into baseNPs and non-baseNPs (abbreviated as ~baseNP). Table-1 gives some examples.

Table-1: Examples of baseNP and ~baseNP

Type	Examples
BaseNP	空中/air 走廊/corridor
BaseNP	政治/politics 体制/system 改革/reform
BaseNP	出口/export 商品/commodity 价格/price 指数/index
~ baseNP	复杂/complicated 的/de 特征/feature
~ baseNP	研究/research 与/and 发展/development
~ baseNP	老师/teacher 写/write 的/de 评语/comment

Both baseNP recognition and baseNP structural analysis are basic tasks in Chinese information retrieval. The paper mainly discusses the problems in structural analysis of baseNPs, which is essential for generating the compositional indexing units from a baseNP. The task of baseNP

* The research is supported by the key project of the National Natural Science Foundation

structural analysis is to determine the syntactic structure of a baseNP. In this paper, we use dichotomy for baseNP analysis. For example, the structure of "自然/natural 语言/language 处理/process" is "(自然/natural 语言/language) 处理/process". Obviously, a baseNP composed of three or more than three words has syntactic ambiguities. For example, baseNP "*x y z*" has two possible structures, that is "(*x y*) *z*" and "*x* (*y z*)". The task of baseNP structural analysis is to select the correct structure from the possible structures.

The paper mainly discusses the problems related to Chinese baseNP structural analysis. Section 2 puts forward a quasi-dependency model for structure analysis of Chinese baseNPs. Section 3 gives an unsupervised quasi-dependency-strength estimation algorithm based on the minimum description length (MDL) principle. Section 4 analyzes the performance of the proposed model and the algorithm. Section 5 discusses some issues in the implementation of baseNP structure analysis and quasi-dependency-strength estimation. Section 6 is the conclusion.

2. The quasi-dependency model

There are two kinds of structural analysis models for English noun phrase, that is adjacency model and dependency model. The research of Lauer shows that the dependency model is superior to the adjacency model for structural analysis of English noun phrase[2]. However, there is no model for structural analysis of Chinese baseNP till now.

According to the dependency grammar, two constituents can be bound together they are determined to be dependent. The determination of the dependency relation between two constituents is composed of two steps. The first step is to determine whether they have the possibility to constituent dependency relation. The second step is to determine whether they have dependency relation in the given context. The former is called the quasi-dependency-relation, which can be acquired from collocation dictionaries or corpora. The determination of the latter is difficult, because multiple information in the given context should be taken into consideration, such as syntax or semantics information, etc.

【Definition 2】 Quasi-Dependency-Relation: If two words *x* and *y* have the possibility to constituent dependency relation, then we say that they have quasi-dependency-relation in the given baseNP, formulated as $x \rightarrow y$ (where *y* is called the head) or $y \rightarrow x$ (where *x* is called the head); Otherwise, we say that they have no quasi dependency relation, formulated as $x \not\rightarrow y$ and $y \not\rightarrow x$.

【Assumption 1】 In a Chinese baseNP, if two words *x* and y can constituent dependency relation, then the head is always the post-positon word *y*, that is $x \rightarrow y$.

According to the Definition 1, there is no preposition phrase, verb phrase, locality phrase or (的)-structure in a baseNP, so assumption-1 is reasonable.

On the basis of assumption-1, we put forward the quasi-dependency model for structural analysis of Chinese baseNPs.

There are the following 3 kinds of quasi-dependency-pattern for a tri-word-composed baseNP *xyz*.

	y	*z*
x	√	×
y	s_{31}	√

s_{31} =(*x y*) *z*

	y	*z*
x	×	√
y	s_{32}	√

s_{32}=*x* (*y z*)

	x	*z*
x	√	√
y	s_{33}	√

Where, pattern s_{31} means $x \rightarrow y$, $y \rightarrow z$ and $x \not\rightarrow z$, which corresponds to structure (*x y*) *z*; pattern s_{32} means $x \rightarrow z$, $y \rightarrow z$ and $x \not\rightarrow y$, which corresponds to the structure *x* (*y z*); However, the quasi-dependency-strength must be used to determine the corresponding structure for pattern s_{33}, which means $x \rightarrow y$, $y \rightarrow z$ and $x \rightarrow z$. For example, as for baseNP "政治/politics 体制/system 改革/reform", there are quasi-dependency-relations "政治/politics→体制/system", "政治/politics→改革/reform" and "体制/system→改革/reform". If we know that the quasi-dependency-relations "政治/politics→体制/system" and "体制/system→改革/reform" are stronger than "政治/politics→改革/reform", the structure of the baseNP can be determined to "(政治/politics 体制/system) 改革

/reform".

In the following, we give the definition of quasi-dependency-strength and the formula for determining the syntactic structure of baseNPs based on the quasi-dependency-strengths.

【Definition 3】quasi-dependency-strength: Given a baseNP set NP=$\{np_1,np_2,...,np_M\}$ and lexicon $W=\{w_1,...,w_M\}$, $\forall w_i, w_j \in W$, the quasi-dependency-strength of $w_i \rightarrow w_j$ is defined as:

$$ds(w_i \rightarrow w_j) = \frac{\sum_{np_k \in NP} dep(w_i \rightarrow w_j, np_k)}{\sum_{np_k \in NP} co(w_i \rightarrow w_j, np_k)}$$

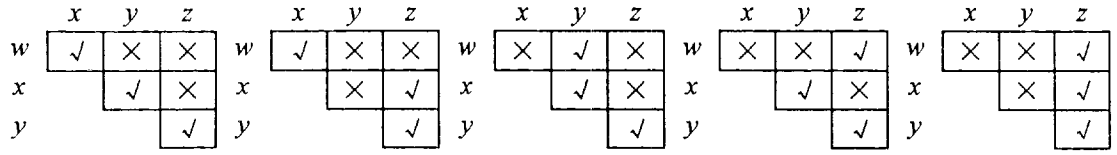
where $dep(w_i \rightarrow w_j, np_k)$ is the count of dependent word pair $w_i \rightarrow w_j$ contained in np_k, $co(w_i, w_j, np_k)$ is the count of cooccurent word pair (w_i, w_j) contained in np_k.

The formula for determining the syntactic structure of baseNP based on the quasi-dependency-strengths is as follows.

$$belief(s_j \mid np_i) = \frac{\sum_{(u \rightarrow v) \in D(np_i, s_j)} ds(u \rightarrow v)}{\sum_{(u \rightarrow v) \in D(np_i, s_j)} ds(u \rightarrow v) + \sum_{(u \rightarrow v) \notin D(np_i, s_j)} ds(u \rightarrow v)}$$

Where, $belief(s_j \mid np_i)$ represents the belief in which the structure of np_i is s_j. $D(np_i,s_j)$ represents the set of quasi-dependency-relations included in the quasi-dependency-pattern corresponding to structure s_j.

A tri-word-composed baseNP has two possible syntactic structures, that is s_{31} and s_{32}. Similarly, a four-word-composed baseNP has the following five possible structures.



$s_{41} = ((wx)y)z$ $s_{42} = (wx)(yz)$ $s_{43} = (w(xy))z$ $s_{44} = w((xy)z)$ $s_{45} = w(x(yz))$

In summary, we can compute the belief in which the structure of np_i is s_j using the correspondence between the quasi-dependency-pattern and the baseNP structure. The acquisition of quasi-dependency-strength between words is the critical problem.

3. The acquisition of quasi-dependency-strength between words

If we have a large scale baseNP annotated corpus in which the baseNPs have been assigned the syntactic structures, the quasi-dependency-strength between words can be acquired through a simple statistics. However, such an annotated corpus is not available. We only have a baseNP corpus which has no structural information. How to acquire the quasi-dependency-strength from such a corpus is the main task of the section. Given a baseNP set $NP=\{np_1,np_2,...,np_M\}$ and a lexicon $W=\{w_1,w_2,...,w_M\}$, the problem can be described as learning a quasi-dependency-strength set G (abbreviated as model) from the training set. Where, $G=\{ds_{ij} \mid ds_{ij}=ds(w_i \rightarrow w_j), \forall w_i, w_j \in W\}$

Zhai Chengxiang puts forward an unsupervised algorithm for acquiring quasi-dependency-strength from noun phrase set[3]. The algorithm is derived from the EM algorithm. Because the algorithm is based on the maximum likelihood (ML) principle, it usually leads to overfitness between the data and the model[4]. For example, given a simple baseNP set NP={政治/politics 体制/system 改革/reform, 经济/economics 体制/system 改革/reform, 政治/politics 体制/system 革命/revolute, 经济/economics 体制/system 革命/revolute}, there are sixteen possible models for the training set, among them G_4, G_7, G_{10} and G_{13} have the best fitness to NP, that is $Num(NP|G)=6$. However, in the linguistic view, G_1 is the correct model, though it has lower fitness to NP, that is $Num(NP|G)=4$ (see the appendix).

3.1 The estimation of the quasi-dependency-strength under Bayesian framework

In Bayesian framework, the task of acquiring the quasi-dependency-strength can be described as the problem of selecting G which has the highest

posterior probability $p(G|NP)$.

$$G = \arg\max_G p(G|NP)$$

According to Bayesian theorem, we have the following inference.

$$G = \arg\max_G \frac{p(NP|G)p(G)}{p(NP)}$$
$$= \arg\max_G p(NP|G)p(G)$$

Besides using conditional probability $p(NP|G)$ to measure the fitness between the training set and the model G, Bayesian modeling gives additional consideration to the generality of the model through the prior probability $p(G)$, that is simpler model has higher probability. The central idea of Bayesian modeling is to find a compromise between the goodness of fit and the simplicity of the model.

3.2 Defining the evaluation function of Bayesian modeling using MDL principle

The difficulty in Bayesian modeling is the estimation of the prior probability $p(G)$. According to the coding theory, the lower bound of the coding length (bit-string) of an information with probability p is $\log_2 1/p$ [5]. The theorem connects Bayesian modeling with the MDL principle in the coding theory.

$$G = \arg\max_G p(NP|G)p(G)$$
$$= \arg\min_G \{-\log_2[p(NP|G)p(G)]\}$$
$$= \arg\min_G \{\log_2 \frac{1}{p(NP|G)} + \log_2 \frac{1}{p(G)}\}$$
$$= \arg\min_G \{L(NP|G) + L(G)\}$$

Where, $L(\alpha)$ is the optimal coding length of information α. Specially, $L(NP|G)$ is called the data description length and $L(G)$ is called the model description length.

Therefore, the problem of estimating the prior probability $p(G)$ and the conditional probability $p(NP|G)$ is converted to the problem of estimating the model description length $L(G)$ and the data description length $L(NP|G)$.

3.3 The MDL-based quasi-dependency-strength estimation algorithm

In MDL principle, the modeling problem can be viewed as a problem of finding a model G which has the smallest sum of the data description length and the model description length. Because the search space is huge, we can not find the optimal model in a transversal manner. The model must be improved in an iterative manner in order to arrive at a minimum description length.

In the research, the model is composed of the quasi-dependency-strength $ds(w_i \rightarrow w_j)$, where each $ds(w_i \rightarrow w_j)$ can be decomposed into two parts: ①the structure part: the quasi-dependency-relation $(w_i \rightarrow w_j)$; ②the parameter part: the quasi-dependency-strength ds. Therefore, the learning process is divided into two steps: ① Keeping the structure part fixed, optimize the parameter part; ②Keeping the parameter part fixed, optimize the structure part. The two steps go on alternately until the process arrives at a convergent point.

Algorithm 1: The MDL-based algorithm for quasi-dependency-strength estimation

---

①Initialize model G;

②Let $L = L(NP|G) + L(G)$, $G = (G_S, G_p)$, where G_S and G_p represent respectively the structure part and the parameter part. Execute the following two steps alternately, until L converged.

- Keeping G_S fixed, optimize G_p, until $L(NP|G)$ converges, that is L converges;
- Keeping G_p fixed, optimize G_s, until $L(G)$ converges, that is L converges.

On condition that the structure part of the model is fixed, the parameter optimization means to find the optimal sets of quasi-dependency-strength in order that the data description length minimized, that is

$$G = \arg\min_G L(NP|G)$$

Where $L(NP|G)$ is the optimal coding length of NP

when G is known.

The parameter optimization step can be implemented using EM algorithm[3]. In the process of parameter optimization, the structure part of the model is kept fixed. The optimum estimates of the parameters are obtained through the gradual reduction of data description length.

In MDL principle, the model description length can be gradually reduced through the modification of the structure part of the model, therefore the overall description length of the model is reduced.

Algorithm 2: The structure optimization algorithm

Let the model after the parameter optimization process is G, which is composed of the quasi-dependency-strength $ds(w_i \rightarrow w_j)$.

①Sort the quasi-dependency-strengths of model G in ascending order, that is $ds^{[1]}$, $ds^{[2]}$, $ds^{[3]}$, ... ;

②Repeat the following steps, until $[L(NP \mid G') + L(G')] - [L(NP \mid G) + L(G)] <= Th_L$ (Th_L is the selected threshold). Let i=1,

- Delete the quasi-dependency-strength $ds^{[i]}$ from model G;
- Construct the new model G';
- **If** $[L(NP \mid G') + L(G')] - [L(NP \mid G) + L(G)] <= Th_L$ **Then** the cycle ends **Else** let $G = G'$, i=i+1 and continue the next cycle.

---

4. The performance analysis

This section takes the N2+N2+N2-type (where N2 represents bi-syllable noun) baseNPs as the testing data in order to discuss the performance of the quasi-dependency-based model for structural analysis of baseNPs and the MDL-based algorithm for quasi-dependency-strength acquisition. The training set includes 7,500 N2+N2+N2-type baseNPs. The close testing set is the 500 baseNPs included in the training set. The open testing set is the 500 baseNPs outside the training set. The testing target is the precision of baseNP structural analysis, that is

$$precision = \frac{a}{b} \times 100\%;$$

Where a is the count of the baseNPs which are correctly analyzed, b is the count of the baseNPs in the tesing set.

4.1 The performance of the quasi-dependency model

The experiments shows: ①In the N2+N2+N2-type baseNPs, the left-binding structure is about two times of the right-binding structure; ②The analysis precision of the quasi-dependency model is about 7% higher than that of the adjacency model. This conclusion can be explained intuitively through the following example. The structure of baseNP "博士/doctor 论文/dissertation 提纲/outline" can not be correctly determined through the adjacency model, because we can not find that the dependency strength of "博士/doctor 论文/dissertation" is stronger than that of "论文/dissertation 提纲/outline". In the other hand, the structure of the above baseNP can be determined to "(博士/doctor 论文/dissertation) 提纲/outline" through the quasi-dependency model, because both "博士/doctor 论文/dissertation" and "论文/dissertation 提纲/outline" are dependent word pairs, while "博士/doctor 提纲/outline" is an independent word pair. Table 2 is the testing result.

Table-2: The analysis precision of N2+N2+N2-type baseNP

Testing type	Right-binding	Left-binding	Adjacency model	Quasi-dependency model
Close test	31.5%	68.5%	84.6%	91.5%
Open test	32.7%	67.3%	81.5%	88.7%

4.2 The performance of the MDL-based algorithm for quasi-dependency-strength acquisition

The ML algorithm is equivalent to the first parameter optimization process of the MDL algorithm. The MDL process is composed of two iterative optimization steps. In the iterative process, the parameters are optimized gradually and the model is simplified gradually as well. Therefore, the overfitness problem inherent in the ML algorithm is solved to a great extent. In the following, the performance of the ML algorithm and the MDL algorithm are compared through comparing the baseNP analysis precision of the models constructed using the above two algorithms. The precision is listed in Table-3. The experiment shows that the MDL algorithm is superior to the ML algorithm.

Table-3: The performance of ML algorithm and MDL algorithm

BaseNP analysis precision			
Close test		Open test	
ML algorithm	MDL algorithm	ML algorithm	MDL algorithm
89.0%	91.5%	82.5%	88.7%

5. Implementation issues

The most difficult problem related to the structural analysis of baseNPs is the acquisition of the quasi-dependency-strength. The proposed algorithm(Algorithm 2) is an unsupervised algorithm, that is the parameters are estimated over the baseNP corpus which has no structural information. In order to improve the estimation results and speed up the iteration process, some measures are taken during the implementation.

5.1 The pre-assignment of the baseNP structure

The structures of some baseNPs can be determined using the linguistic knowledge. Such knowledge includes:

① In a baseNP, a word pair which has the following syntactic composition is independent.

- Noun+Adjective: for example, "**地面/ground/Noun 复杂/complicated/Adjective** 状态/condition", "**玻璃/glass/Noun 弯/curved/Adjective** 管/pipe";
- Noun+Distinctive: for example, "**小学/elementary-school/Noun 适龄/of-the-right-age/Distinctive** 儿童/child";
- Distinctive+Verb: for example, "**大型/large/Distinctive 作战/fight/Verb** 飞机/plane", "**低级/elementary/Distinctive 爬行/creep/Verb** 动物/animal".

② If two verbs cooccur in a baseNP, then they are dependent. For example, "（**勘察/prospect/Verb 设计/design/Verb**）单位/group", "（**抗日/Anti-Japanese/Verb 救国/save-the-nation/Verb**）运动/campaign".

If we preprocess the baseNP corpus using the above knowledge, it is beneficial for the estimation process.

5.2 The complex-feature-based modeling

If the lexicon size is $|W|$, then the parameter number of the above word-based acquisition algorithm amounts to $|W|^2$. The enormous parameter space will lead to the data sparseness problem during the estimation. Therefore, the paper puts forward the complex-feature-based acquisition algorithm. First, map each word to a complex-feature-set according to the multiple feature of the words; Then, acquire the quasi-dependency-strength between the complex-feature-sets. During analyzing the structure of a baseNP, the strength between the complex-feature-sets is used instead of that between the words. In the research, the multiple features include part-of-speech, number of syllables and word sense categories.

6. Conclusions

The paper put forward a quasi-dependency model for structural analysis of Chinese baseNPs, and a MDL-based algorithm for the quasi-dependency-strength acquisition. The experiments show that the proposed model is more suitable for Chinese baseNP analysis and the proposed MDL-based algorithm is superior to the traditional ML-based algorithm. The further research will focus on incorporating more linguistic knowledge into the above statistical model.

References

[1] Church K., A stochastic parts program and noun phrase parser for unrestricted text, In: Proceedings of the Second Conference on Applied Natural Language Processing, 1988.

[2] Lauer M. Conceptual association for compound noun analysis, In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Student Session, Las Cruces, NM, 1994.

[3] Zhai Chengxiang, Fast Statistical Parsing of Noun Phrases for Document Indexing, In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, USA.: Association for Computational Linguistics. 1997. 311-318.

[4] Stolcke A. Bayesian learning of probabilistic language models, Dissertation for Ph.D. Degree, Berkeley, California: University of California, 1994.

[5] Solomonoff R. The mechanization of linguistic learning, In: Proceedings of the 2nd International Conference on Cybernetics.

Appendix: An example for quasi-dependency-relation acquisition

No.	Model G	|G|	Fitness between G and NP	Num(NP|G)
1	政治体制(1), 经济体制(1) 体制改革(1), 体制革命(1) 政治改革(0), 经济改革(0) 政治革命(0), 经济革命(0)	4	(政治体制)改革(1), 政治(体制改革) (0) (经济体制)改革(1), 经济(体制改革)(0) (政治体制)革命(1), 政治(体制革命) (0) (经济体制)革命(1), 经济(体制革命)(0)	4
4	政治体制(1), 经济体制(1) 体制改革(1), 体制革命(1) 政治改革(0), 经济改革(0) 政治革命(1), 经济革命(1)	6	(政治体制)改革(1), 政治(体制改革) (0) (经济体制)改革(1), 经济(体制改革)(0) (政治体制)革命(1), 政治(体制革命) (1) (经济体制)革命(1), 经济(体制革命)(1)	6
7	政治体制(1), 经济体制(1) 体制改革(1), 体制革命(1) 政治改革(0), 经济改革(1) 政治革命(1), 经济革命(0)	6	(政治体制)改革(1), 政治(体制改革) (0) (经济体制)改革(1), 经济(体制改革)(1) (政治体制)革命(1), 政治(体制革命) (1) (经济体制)革命(1), 经济(体制革命)(0)	6
10	政治体制(1), 经济体制(1) 体制改革(1), 体制革命(1) 政治改革(1), 经济改革(0) 政治革命(0), 经济革命(1)	6	(政治体制)改革(1), 政治(体制改革) (1) (经济体制)改革(1), 经济(体制改革)(0) (政治体制)革命(1), 政治(体制革命) (0) (经济体制)革命(1), 经济(体制革命)(1)	6
13	政治体制(1), 经济体制(1) 体制改革(1), 体制革命(1) 政治改革(1), 经济改革(1) 政治革命(0), 经济革命(0)	6	(政治体制)改革(1), 政治(体制改革) (1) (经济体制)改革(1), 经济(体制改革)(0) (政治体制)革命(1), 政治(体制革命) (1) (经济体制)革命(1), 经济(体制革命)(0)	6