A method for Chinese Noun Phrase Recognition Base on Word Co-occurrence Directed Graph

Xinglin Liu

School of Computer Science, Wuyi University, Jiangmen, 529020, China
jmxlliu@163.com

Abstract—**This paper proposes a recognition method for Chinese Noun Phrase based on word co-occurrence directed graph. An input document is firstly scanned in which noun word string is retrieved. Atomic word table and word co-occurrence directed graph is then generated according to the word strings. A search is performed on the graph to find the longest paths with priority weight satisfying certain criteria. The word strings corresponding to the paths are considered as noun phrases. As dimensionality reduction is applied, the scale of the word co-occurrence directed graph is reduced significantly, and thus the efficiency of the algorithm is improved. Experimental results demonstrate that the precision of noun phrase recognition reaches 95.4%.**

Keywords-Atomic word; Noun phrase; Word co-occurrence directed graph; Knowledge acquisition; Natural Language Processing

I. INTRODUCTION

The Chinese distinguishing word is a fundamental research topic in the area of Natural Language Processing. The correctness of Chinese distinguishing word directly affects the subsequent processes. Currently, mainstream distinguishing word systems are based on string alignment, which is dependent on dictionaries. Consequently, when the text to be processed contains words that are not included in the dictionary, the words will be recognized as multiple separate words or morphemes. Noun phrase recognition has its practical values in the area of Machine Learning, Text Information Retrieval, and Information Extraction, etc.

YANG[1] proposed the methodology of classifying words into two categories, i.e., atomic words and compound words. An atomic word refers to a morpheme or sub-word sliced and recognized by the distinguishing system, and will constitute other new words; and compound words have complete meanings that are consisted of atomic words. As most words in a document are compound words including Noun Phrases. The reorganization of such compound words becomes fundamental and meaningful.

Based on intensive statistical analysis, YUAN et al [2] announced that the percentage that the meaning of a compound word was the composition of the meanings of its morphemes accounts for 87.8%, 93.2% and 87.0%, in nouns, verbs, and adjectives, respectively. More specifically, among Noun Phrases, words containing two noun short words account for 46.7%, and 57.2% for that of two noun morphemes.

In this paper, we are concerned about Noun Phrases which have complete meanings and are consisting of $L(L \geq 2)$ noun atomic words. According to the analytical results in [2], Noun Phrases following the form of "noun+…+noun" take a large percentage, without changing the meanings of the components. CHEN et al [3] concluded that an important word would appear multiple times in a document, with different words ahead of and after it. Based on the two points mentioned above, we propose a recognition method for Chinese Noun Phrase based on word co-occurrence directed graph, in which word co-occurrence means that multiple successive atomic words appear in the same sentence without any symbols between them. These atomic words form a word string. The word co-occurrence directed graph is then generated according to these word strings, with atomic words in the word string as the vertices in the graph, and edges connecting these co-occurrence vertices.

The remainder of this paper is organized as follows. We first give a brief survey on existing work related to this paper in Section II. We detail our proposed Noun Phrase recognition method in Section III. We present the experimental results on ten 863-Evaluation-Corpus-Documents in Section IV. Finally, conclusions are given in Section V.

II. RELATED WORK

Borrowing the idea of the cognitive mental model of human, CHEN et al [3] proposed a new detection algorithm based on directed net of word-sequence frequency to discover combined-word, for the recognition of combined-word in free texts. In the algorithm, a directed net characterizing the frequency of occurrence of word-sequence in a text document is first established. Then a specific matrix operation is performed to retrieve combined-words step by step. The advantages of the algorithm lie in little dependence on the linguistic expertise, and the precision can reach 90.2% demonstrated by their experimental results.

In 《Chinese BaseNP Recognition and Structure Analysis》, ZHAO [4] proposed the formularized definition for Base Noun Phrase (BaseNP here after), which was the first work in China on the recognition of BaseNP.

ZHANG et al [5] developed an analytical model for the BaseNP of Chinese language based on knowledge graphs. The Chinese knowledge graphs is created using 《HowNet》 as its source of the semantic knowledge. Semantic is of their main concern, with syntax as the

secondary consideration. "Knowledge graphs" are constructed by making sentences for every notional, and these "Knowledge graphs" are then combined together to create a "Phrase graphs", which eventually leads to a knowledge graphs including the structural information of the BaseNP and the semantic information in the Chinese language. The precision of the BaseNP recognition can reach 83.6%.

XU et al [6] adopted a combined classification method based on Error-driven to recognize Chinese BaseNP. The approach is integrated with CRF, SVM and TBL (Transformation Based Learning) model , and applies the SVM model to learn the rules of wrongly-processed corpuses by CRF and TBL. By making use of multiple classifiers, the recognition precision can be improved to 89.72%。

Concerning recognition of unregistered words, ZHOU [7] proposed a statistic and rules-based method. The input text is firstly segmented and tagged with Part-of-Speech for each word, while a temp dictionary is generated. Each word string is granted a weight according to the rule and its occurrence frequency. Then a greedy algorithm is used to find the longest path for each fragment, and thus the unregistered word can be then recognized. Experiments show that the precision can reach 81.25%.

LI [8] proposed a recognition method of the Hidden Markov Model (HMM) based on the genetic algorithm, which is developed on the basis of the high correctness of the Part-of-Speech Tagging. During the training stage, genetic algorithm is applied to gain HMM parameters, and in the recognition stage, an improved Viterbi algorithm is used to perform dynamic programming, and recognize Noun Phrase in the same layer. Then the hierarchical syntax parsing algorithm and the improved Viterbi algorithm are combined together to recognize recursive Noun Phrases. The combined algorithm can achieve a precision as 94.78, and the recall is 94.29%.

Church [9] used the boundary statistics method to recognize BaseNP in the English language. In this method, it starts to count the Part-of-Speech symbol at the beginning and ending positions of Noun Phrases from tagged corpus, which will result in two matrices: the probability matrix for the beginning positions of the Noun Phrases, and that of the ending positions. Then it gets the two neighboring Part-of-Speech Tag from an input sentence, retrieves the largest probability from the probability matrix, then marks the start and the end, and thus recognizes BaseNP. However, the paper did not present experiment results.

Erik et al. [10] generated different classifiers by using different representations of the data. By combining the results with voting techniques, they managed to improve the best reported performances on standard data sets for baseNPs and arbitrary Noun Phrases.

Taku et at [11] adopted Support Vector Machines (SVMs) to identify English base phrases (chunks). The SVMs can summarize the multi-dimension input data. Besides, SVMs are able to finish training with little computing cost by using the properties of multi-dimension space. During the post-processing phase, to further improve the recognition ability, a combined system based on weighted voting mechanism is constructed, which is composed of 8 SVMs models by training on different combination of features. The recognition precision can reach up to 95.77%.

III. NOUN PHRASE RECOGNITION ALGORITHM

We consider a word string (composed of $L(L \geq 2)$ successive atomic words in a sentence) as a Noun Phrase if it satisfies following conditions.

a) All atomic words in the word string are nouns.
b) The word string has multiple occurrences in the entire document.
c) Such case is rare that a noun atomic word is in front of or after the word string, and the atomic word and the word string form a new word string.

For future convenience, we refer the aforementioned determination to Hypothesis-1. Then we present a text section that is already segmented and tagged with Part-of-Speech for each word. The example (Sample-1 here after) will be used throughout the paper.

① *In today, information system offers for enterprise communications and analysis capabilities, so that it can guide trade and management business worldwide.* ②*Global enterprise communication with distributors and suppliers, under the environment of different countries operate 24 hours a day, and provide services to a range of local and international demand, therefore, to control the global enterprise is a major business challenge, it needs a powerful information system response.*

In Sample-1, ① and ② represent the ID of the two sentences, which are for the purpose for presentation and do not exist in the tagged text. The work in this paper is based on the three conditions of the Noun Phrase mentioned above.

A. Definition for Word co-occurrence Directed Graph

A word co-occurrence directed graph is denoted as $G :< V, E >$, where V refers to the set of all atomic words in the document, and E is the set of the word pairs. The starting point of an edge is the head-word of the word pair, and the ending point is the tail-word. The weight of each edge is a set of the positions of the occurrence of word pairs in the document, with a triple $< sno, start, end >$ as the element, which means the sentence ID, the starting position and end position of the word pair in the sentence.

We explain the terminologies that will be used later in the paper. In V, the elements are $v_1, v_2, ..., v_{|V|}$, e_{ij} is a directed edge with v_i and v_j as its starting point and ending point, respectively. s_{ij} is the set of the triple $< sno, start, end >$ on the edge, and $w_{ij} = |s_{ij}|$ is the value of the weight. $p < v_i, ..., v_j >$ represents the path of the corresponding vertices $v_i, ..., v_j$. $ps < v_i, ..., v_j >$ is the intersection set of all the edge sets of the corresponding

path, which is also referred to as the set of $p < v_i,...,v_j >$. $len < v_i,...,v_j >$ means the length of path $p < v_i,...,v_j >$, and $weight < v_i,...,v_j >= \left| ps < v_i,...,v_j > \right|$ represents the value of the weight of $p < v_i,...,v_j >$. We define a analogous intersection operator $\cap^s$ as the set intersection operation for the edges of the word co-occurrence directed graph as follows.

$$X \cap^s Y = \{< sno, start, end > | < sno, start, mid > \in X, < sno, mid, end > \in Y\} \quad (1)$$

It can be seen that $X \cap^s Y \neq Y \cap^s X$. Therefore, when performing the analogous intersection operation, the ending point of the edge (or path) of the left operand must be the starting point of the right operand.

As explained above, in the word co-occurrence directed graph, if the corresponding word string of a path $p < v_i,...,v_j >$ satisfies the three conditions of Hypothesis-1, the word string will be recognized as a Noun Phrase.

B. Generates Word co-occurrence Directed Graph

Usually, the scale of the set of atomic word in a document is very large. In [12], Heaps found that the scale of a Corpus was related to the capacity of vocabulary as follows:

$$v = k \cdot n^{\beta} \quad (2)$$

where v is the capacity of the vocabulary, n is the number of words in the corpus, k and β are parameters according to different corpus, which satisfies $10 \le k \le 100$ and $\beta \approx 0.5$. We set $k = 60, \beta = 0.5$. Then in a document including 6000 words, the capacity of the vocabulary is around 4647. Obviously, the scale of the word co-occurrence directed graph generated from the vocabulary is very large.

To reduce the scale of the graph, we retrieve noun word string using condition a) when scanning the document. For example, table I shows the noun word strings after Sample-1 is scanned.

TABLE I. NOUN WORD STRING OF SAMPLE-1

Noun Word Strings	Position
information system	<1,4,5>
global range	<1,19,20>
global enterprise	<2,1,2>
international range	<2,27,28>
global enterprise	<2,37,38>
information system	<2,51,52>

From table I we can see that $\left|V\right| = 7$, which is much smaller, compared with the scale as 54 of the vocabulary in Sample-1.

The word co-occurrence directed graph generated using table 1 is plotted in Fig. 1. It can be seen that the scale of the graph has been significantly reduced, which will relieve the computational cost in the following steps.

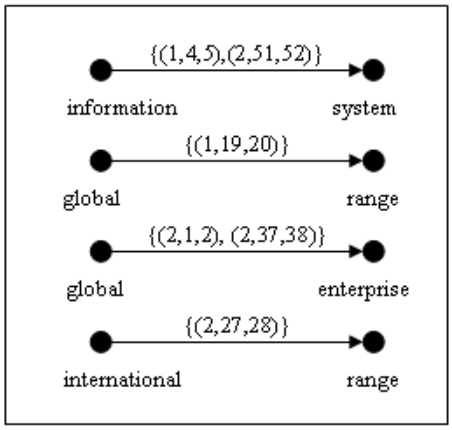

Figure 1. Word co-occurrence Directed Graph of Sample-1

A word co-occurrence directed graph is related to an adjacency matrix, which is denoted as M, with each element as a set, i.e., s_{ij}. The corresponding matrix of Fig. 1 is shown in Fig. 2. (For future convenience, we show the matrix as a two-dimensional table, with added row title and column title.)

	information	system	global range	global	enterprise	international
information		{<1,4,5>, <2,51,52>}				
system						
global range			{<1,19,20>}			
global					{<2,1,2>, <2,37,38>}	
enterprise						
international			{<2,27,28>}			

Figure 2. Matrix for Figure 1

In Fig. 2, M is a sparse matrix. Each row represents the head-words of the word pairs, and a column shows the tail-words. If $m_{ij} \neq \Phi$, there exists a directed edge from v_i to v_j.

C. Noun Phrase Recognition

In our algorithm, we set an integer threshold T. If $weight < v_i,...,v_j >= \left| ps < v_i,...,v_j > \right| < T$, the word string is considered not being a Noun Phrase; if $weight < v_i,...,v_j >= \left| ps < v_i,...,v_j > \right| \ge T$ and $weight < v_{i-1} v_i,...,v_j >= \left| ps < v_{i-1} v_i,...,v_j > \right| < T$, the word string corresponding to $p < v_i,...,v_j >$ is treated as a Noun Phrase.

In general cases, we set $T = 2$. Using this value, Fig. 1 can be simplified by deleting edges that satisfying

$w_{ij} = |s_{ij}| < T$ and removing isolated vertices. We refer to such process as dimensionality reduction. By dimensionality reduction, Fig. 1 can be reduced to Fig. 3, with its adjacency matrix as Fig. 4.

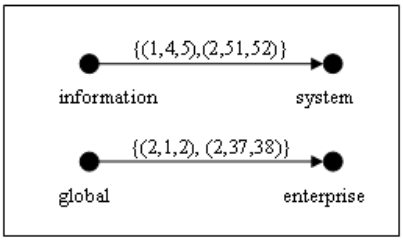

Figure 3. Dimension Reduced Word co-occurrence Directed Graph of Sample-1

	system	enterprise
information	{<1,4,5>, <2,51,52>}	
global		{<2,1,2>, <2,37,38>}

Figure 4. Matrix for Figure 3

We borrow the idea of the Bellman-Ford algorithm and design an algorithm to find the longest path with largest weight value for multiple starting points in the word co-occurrence directed graph. We use the algorithm to find Noun Phrases.

The algorithm is presented as follows.

1) Set $s = aMatrix(1,0)$, d is Null, $path = s$, $len(path) = 1$, $ps(path) = \Phi$, $weight(path) = 0$.
2) Set s to be the source node, searching the next node d in the word co-occurrence directed graph. If fails, go to step 3. Otherwise, do step a) to d).
   a) Calculates $ps(path) = ps(path) \cap^{s} ps < s, d >$.
   b) Updates $weight(path)$.
   c) If $weight(path) \geq T$, then set $path = path \& d$, $len(path) += 1$, set $s = d$, d is Null.
   d) Turning to step 2.
3) If $len(path) < L$, go to step 7.
4) Store the recognized Noun Phrase.
5) Delete the extracted path of the Noun Phrase from the word co-occurrence directed graph.
6) Perform dimensionality reduction to the word co-occurrence directed graph.
7) If the graph is not null, then return to step 1.
8) Output all recognized noun phrase, and algorithm terminates.

Through the iterative search of the algorithm, it is able to find the longest path satisfying the 3 conditions of Hypothesis-1, and thus the Noun Phrases are recognized.

As the length-first strategy is applied, longer Noun Phrase will be recognized first. The benefit lies in the fact if a Noun Phrase contains another Noun Phrase, these two Noun Phrases can be recognized in order.

IV. EXPERIMENTS

We select 10 863-Evaluation-Corpus Documents as our experiment data set. The average number of words is around 3017. And ICTCLAS developed by Chinese Academy of Science is used as the distinguishing word system.

To evaluate the correctness of recognizing a Noun Phrase, if the Noun Phrase indeed represents a compete meaning in the document, the recognition is considered as 'correct'. Table II shows the experimental results.

TABLE II. THE RESULT OF EXPERIMENTS

Noun Phrase Recognized Number	Right Number	Precision
7	7	1.0000
5	5	1.0000
10	9	0.9000
15	13	0.8667
10	9	0.9000
23	23	1.0000
13	13	1.0000
14	13	0.9286
18	18	1.0000
18	17	0.9444

From table 2, only one document is of the precision ratio lower than 90%, which accounts for 10% of all the documents. Five documents have a precision ratio equal to 100%, taking 50% of the documents. The average precision is 95.40%.

We have studied the Noun Phrases that are not correctly recognized and find that the results, according to our algorithm can only solve the case of "noun+…+noun", words marked as other Part-of-Speech will be filtered out, which leads to incorrect recognition of such Noun Phrase. We will leave the problem to our follow-up work

As we are using a data set different from other strategies, we are not able to conduct the comparison study between other literature and our work, we use the precision mentioned in their papers and our method, and demonstrate the comparison result in Fig. 5.

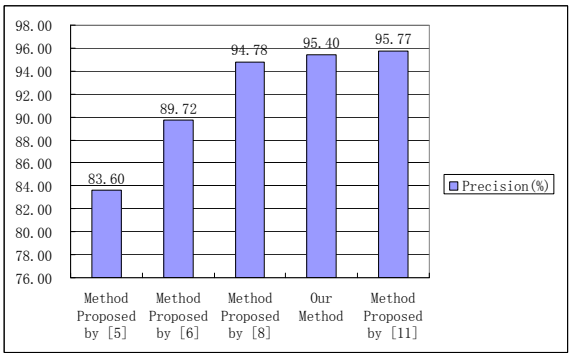

Figure 5. Compare with other methods

Fig. 5 illustrates the efficiency of our algorithm. The precision of our method is only lower than [11] by 0.37%, but higher than other methods in [5][6][8] .

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, the Hypothesis-1 for determining a Noun Phrase is verified by experimental results. A recognition method is proposed based on the word co-occurrence directed graph. It generates the word co-occurrence directed graph according to the noun atomic word strings that it has retrieved. Then a search for the paths satisfying the condition of a Noun Phrase is performed in the directed graph, and thus looks for Noun Phrases. The experiments show that the average correctness ratio can reach as high as 95.40%, which verifies the efficacy and feasibility of our approach. With further improvement, it can be also applied to the recognition of BaseNP and compound words.

As the performance of the recognition relies on the correctness of distinguishing words, in our future work the distinguishing word system will be improved to increase recognition precision. Now our algorithm can only recognize the structure of "noun+…+noun". We will extend it to cover "gerund + noun", "verb + noun" and "adjective + noun" cases. Moreover, the value of the integer threshold T will affect the performance of the recognition of Noun Phrases. Currently a Noun Phrase of weight lower T is not able to be recognized. Our future work will involve combining semantic interpretation in the context to recognize such Noun Phrases.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of Guangdong Province, China (Grant No. S2011010003681); Guangdong science and technology plan projects(Grant No. 2010B010600039); Guangdong Jiangmen science and technology plan projects; Dr. start foundation of Wuyi University(Grant No. 30813008).

REFERENCES

[1] Mei YANG. The construction of modern Chinese compound word. Nanjing: Nanjing Normal University, 2006. (In Chinese)

[2] Chun-fa YUAN, Chang-ning HUANG. "Research of Chinese Morpheme and Word Formation Rules Base on Morpheme Database," Journal of Linguistics Application. No. 3, pp. 83-88, 1998..(In Chinese)

[3] Jian-chao CHEN, Qi-lun ZHENG, Qing-yang LI, Gui-duo YAN. "Chinese combined-word detection based on directed net of word-sequence frequency," Application Research of Computers. Vol. 26, No. 10, pp. 3746-3749, 2009.(In Chinese)

[4] Jun ZHAO. Chinese BaseNP Recognition and Structure Analysis. Beijing : Tsinghua University, 1998.

[5] Rui-xia ZHANG, Lei ZHANG. "The Model for Chinese BaseNP Analysis Based on Knowledge Graphs," Journal of Chinese Information Processing. Vol. 18, No. 3, pp. 47-53, 2004.(In Chinese)

[6] Fang XU, Qing-cheng ZONG, Xia WANG. "Chinese Base NP Chunking by Error-driven Combination Classifiers," Journal of Chinese Information Processing. Vol. 21, No. 1, pp. 115-119, 2007.(In Chinese)

[7] Lei ZHOU, Qian-ming ZHU. "Research on Recognition Method of Unknown Chinese Words Based on Statistic and Regulation," Computer Engineering. Vol. 33, No. 8, pp. 196-198, 2007.(In Chinese)

[8] Rong LI, Jia-heng ZHENG, Mei-ying GUO. "Application Study of Hidden Markov Model Based on Genetic Algorithm in Noun Phrase Identification," Computer Science. Vol. 36, No. 10, pp.244-246, 2009.(In Chinese)

[9] Church K W. A "stochastic parts program and noun phrase parser for unrest ricted text," Proceedings of the 2nd Conference on Applied Natural Language Processing, Texas, USA, 1988, pp.136-143.

[10] Erik F, Sang Tjong Kim. "Noun Phrase Recognition by System Combination," Proceedings of ANLP2NAACL 2000, Seattle, WA, USA, 2000, pp. 50-55.

[11] Taku Kudo and Yuji Matsumoto. "Chunking with Support Vector Machines," Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, USA, 2001.

[12] Heaps, H. Information Retrieval: Computational and theoretical aspects. New York: Academic Press, 1978.