Incorporating Word Clustering Into Complex Noun Phrase Identification*

Xue Lihua[1,*], Zhang Guiping[1], Zhou Qiaoli[1], Ye Na[2]

[1,2]Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang 110136
375618003@qq.com, zgp@ge-soft.com,
zhou_qiao_li@hotmail.com, yena_1@126.com

Abstract. Since the professional technical literature include amounts of complex noun phrases, identifying those phrases has an important practical value for such tasks as machine translation. Through analysis of those phrases in Chinese-English bilingual sentence pairs from the aircraft technical publications, we present an annotation specification based on the existing specification to label those phrases and a method for the complex noun phrase identification. In addition to the basic features including the word and the part-of-speech, we incorporate the word clustering features trained by Brown clustering model and Word Vector Class (WVC) model on a large unlabeled data into the machine learning model. Experimental results indicate that the combination of different word clustering features and basic features can leverage system performance, and improve the F-score by 1.83% in contrast with the method only adding the basic features.

Keywords: complex noun phrase; word clustering; Brown clustering model; word vector class model

1 Introduction

With the development of the aviation industry, translation for aircraft technical publications [1] becomes a significant problem to be solved. Because the sentence structures are simple and the descriptions are mostly on aircraft parts in technical publications, the translation quality of the terminology will affect the translation quality of the whole sentence. Thus, it is necessary to automatically indentify the terminology before acquiring their translations.

Instead of terminology, we use the concept of complex noun phrase with a certain structure to achieve the desired effect. The complex noun phrases should meet some requirements as follows: 1) formed by two or more words according to certain relationship; 2) cannot be included by other noun phrase; 3) not including "的" . There are two reasons for the definition. First, through analysis on the ter-

* This work is supported by Humanities and Social Sciences Foundation for the Youth Scholars of Ministry of Education of China (№-14YJC740126) and National Natural Science Foundation of China (№-61402299).

minology in technical publications, there are high similarities in grammar structures among the large number of terminologies, such as “前 货舱” (the fwd cargo compartment) and “后 货舱” (the aft cargo compartment) in which “前” and “后” are nouns of location. Second, with the rapid development of aviation industry, a considerable number of new terms are created on a regular basis, but the new terms are always consistent in the structure with the old ones. Therefore, considering the structure of the term, we use complex noun phrase, not terminology. For labeling the complex noun phrases, we present a specification based on the bilingual corpus and the existing annotation scheme [2].

To automatically identify complex noun phrases, many supervised learning methods are applied. These methods require sufficient labeled data to achieve state-of-the-art performance. Although supervised learning methods can rapidly develop a robust phrase recognition system, the requirement of substantial amounts of training data is still an impediment to the quick deployment of phrase recognition in new languages or domains. However, it is often the case that developing sizable training data is considerably time-consuming whereas the amount of raw data is rapidly increasing. To further exploit the effects of unlabeled data, several studies investigated on how to incorporate raw data. Koo et al. [3] demonstrated excellent performance on dependency parsing in the use of word clustering feature. Word clustering is a technique for partitioning sets of words into subsets of semantically similar words. For example, “前” in “前 货舱” and “后” in “后 货舱” are similar under this definition, whereas “切断” (cut off) and “开关” (switch), although semantically related, are not. Intuitively, in a good clustering, the words in the same cluster should be similar. And word clustering is increasingly becoming a major technique used in a number of NLP tasks ranging from word sense or structural disambiguation to information retrieval and filtering. So unlike previous work, besides the word and part-of-speech (POS) tags as features for recognizing phrases, we incorporate word clustering [4] feature. We adopt Semi-supervised Learning (SSL) [5,6] techniques to incorporate word cluster into ML model for phrase recognition. SSL is an ML approach that typically uses a large amount of unlabeled data and a small amount of labeled data to build a more accurate classification model than the models using only labeled data. In this paper, besides the word and POS features, we apply word cluster features, trained by Brown clustering [7] model and Word Vector Class (WVC) model on a large amount of unlabeled data, into Conditional Random Fields (CRF) [8].

2 Related work

The noun phrase identification problem can be considered as a sequence labeling problem, whose task is that under the condition of a given observation training x, estimate the conditional probability of sequence. And researchers have done a lot of work on sequence labeling problems like chunk recognition, named entity recognition, word segmentation and part-of-speech tagging and so on.

Work on Chinese sequence labeling problem includes: Sun Ruina [9] combines

the statistics-based and rule-based method, first performing base noun phrase boundary prediction through the mutual information between the words, and then adjusting the boundary prediction for base noun phrase identification according to the constitutive rules of base noun phrases. But this method can't identify the low-frequency phrases in case of sparse corpus. Wang Meng [10] et al adopted a method for automatic acquisition of Chinese compound noun phrases based on corpus, making use of the statistical indexes to get the typical, frequent compound noun phrases as training data and extracting the various features from the training set to help find the infrequent ones. But the method viewed the compound noun phrase acquisition as a static problem and did not use the context information of "Noun-Noun" sequence. Li Guochen [11] et al present a base chunk identification model based on deep neural network models, which takes Chinese characters as tagging unit and original input layer. The results show that the method is useful. But they did not integrate more abundant features of characters like POS, collocation information into the system. Zhang Kaixu [12] adopted a method for a joint Chinese word segmentation and POS tagging task. First, extract high-dimensional distributional lexical information from a large scale unlabeled corpus, then perform unsupervised dimension reduction for the low-dimensional lexicon features by an auto-encoder. Results show that the additional lexicon features improve the performance and are better than those features learned by using the principal component analysis and the k-means algorithm for phrase recognition.

Work on sequence labeling problem of other languages includes: Tsendsuren Munkhdalai [13] et al adopt Semi-Supervised Learning techniques to incorporate domain knowledge into the Chemical and biomedical Named Entity Recognition model, and the results show the method leverage overall system performance. Yu-Chieh Wu [14] presents a cluster-based method to fuse labeled training and unlabeled raw data. They derive the term groups from the unlabeled data and take them as new features for the supervised learners in order to improve the coverage of lexical information. LING ZHU [15] et al present a noun phrase chunking model based on Selection Base Classifiers on Bagging (SBCB) ensemble learning algorithm. The algorithm employs multiple learners and integrates their prediction capabilities to achieve a more accurate classification outcome instead of assembling all classifiers candidates. Results show that the proposed approach is able to achieve a remarkably better performance, which is superior to several comparable state-of-the-art chunking algorithms that apply SVM, HMM, and CRF as well. Michal Konkol et al [16] propose new features for Named Entity Recognition (NER) based on latent semantics and experimented with two sources of semantic information: LDA and semantic spaces. Results show that the newly created NER system is fully language-independent thanks to the unsupervised nature of the proposed features and it achieves the same or even better results than state-of-the-art language-dependent systems.

From the work above, we can see that Semi-Supervised Learning (SSL) techniques have been applied to many NLP tasks such as phrase recognition, word segmentation and POS tagging. And many studies have proved that the features extracted from a large-scale corpus can better reflect the syntactic and semantic

features of words. Thus we incorporate word clustering features trained by Brown clustering model and Word Vector Class (WVC) model on a large amount of unlabeled data respectively.

3 Annotation Specification

Considering the characteristics of complex noun phrases in the Chinese-English bilingual sentence pairs of aircraft technical publications, we refer to the annotation scheme for Chinese Treebank [2, 17] and change some rules of the annotation scheme to identify complex noun phrase. The rules modified are shown in Table 1.

Table 1. Comparison of Tinghua annotation specification and the proposed annotation specification

Phrase structure	Tsinghua	The proposed
np+s	[sp 青尼罗河/n 上游/s]	[np 青尼罗河/n 上游/s]
vp+att-c	进行/v 了/ uA [vp 实地/d 调查/v]	进行/v 了/uA [np 实地/d 调查/v]

np: the nominal phrase; s: the locative word; att-c: Attributive-centered structure

In the following subsections we will introduce the detail descriptions of the rules above.

3.1 np+s

Through analysis of the phrases in bilingual corpus, when s in “np+s” is the words like “上游”, “顶部”, “底部”, “左后侧” et al, its corresponding translation is noun. Thus “np+s” is labeled as complex noun phrase. Some examples about “np+s” Chinese-English annotation and its phrase alignment are shown in Table 2.

Table 2. “np+s” Chinese-English annotation and its phrase alignment

位于 CNP[后设备舱 左后侧] be on CNP[the left rear side of the aft equipment compartment]
从 CNP[APU 引气阀 上游] from CNP[the upstream of the APU]
安装 在 CNP[机身 顶部、底部] Installed on CNP[the top and the bottom of the fuselage]

3.2 vp+att-c

att-c is a phrase structure called attributive-centered structure as a linguistic form of attributives modifying the headword, and it plays an important role [18] both in Chinese and English. Particularly unusual is that the headword is verb acting as noun and the attribute modifying the headword may be adjectives, verbs or nouns. Some examples about "vp+att-c" Chinese-English annotation and its phrase alignment are shown in Table 3.

Table 3. "vp+att-c" Chinese-English annotation and its phrase alignment

用于 CNP[信号 分发] used for CNP[signal distribution]
对 ADC 进行 CNP[转换 选择] perform CNP[conversion selection] to ADC
可以 实现 CNP[减速 控制] CNP[The speed brake control] can be activated
导致 CNP [控制面 非 指令 打开] result in CNP[uncommanded deployment of the control surface]
用以 CNP[逻辑 决断 或 共同 处理] are used for CNP[logic discretion or common processing]

4 Method Description

Our complex noun phrase recognition system design is shown in Fig. 1. First, we perform preprocessing on aircraft technical publications and then extract two different feature sets, a base feature set and a word clustering feature set, in the feature processing phase. The unlabeled data is fed to build word classes. Finally, we apply the CRF sequence-labeling method to the extracted feature vectors to train complex noun phrase recognition model. These steps will be described in subsequent sections.

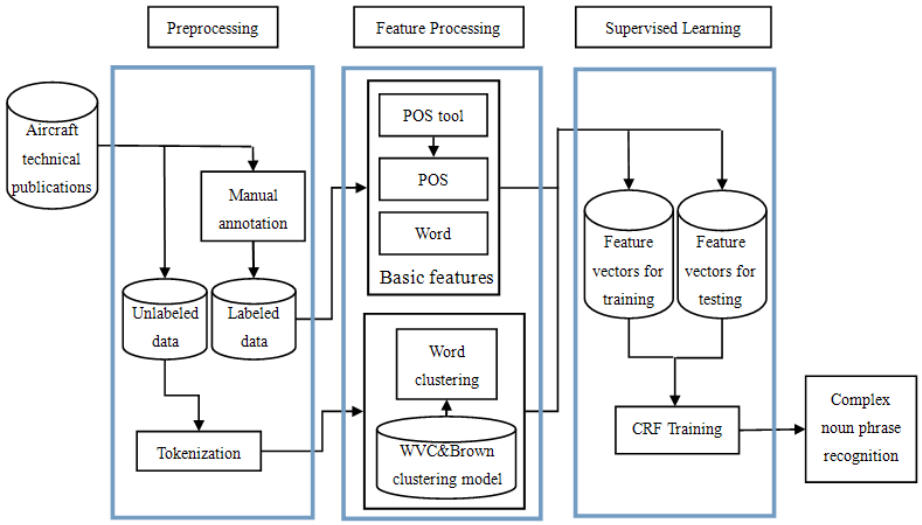

Fig.1. System design for complex noun phrase recognition

4.1 Preprocessing

Preprocessing mainly includes labeling 10 thousand sentences from aircraft technical publications according to the annotation specification modified in section 3. Then the 10 thousand labeled sentences are used as training set for complex noun phrase recognition. Meanwhile, all raw corpora are tokenized by a tokenization tool and used as the input corpus to achieve the word clustering features.

4.2 Feature processing

We extract features from the preprocessed text to represent each token as a feature vector, and then apply them to an ML algorithm for building a model for complex noun phrase recognition.

The proposed method includes extraction of baseline and the word clustering feature sets. The baseline feature set is essential in phrase recognition, but is poor at representing the word clustering feature. In this paper, we achieve the word clustering features by training on a large amount of unlabeled data.

We found that with the increase of window sizes, the performance of recognition becomes better. But due to the limitation of tool for modeling, the window sizes can only be increased to 4. Therefore, when choosing feature templates, we not only consider the current word and its part-of-speech tag, but also consider the long-distance associated words in the previous context, so as to add the richer information mentioned above. The baseline feature set is summarized in Table 4.

Table 4. The baseline features

Features	Feature description
W	The current word
P	POS of W
X	The fifth word before W
P	The sixth word before W

For word clustering features, we train Brown clustering models and word vector class (WVC) models on a large amount unlabeled data. Brown clustering is a bottom-up agglomerative word clustering algorithm to derive a hierarchical clustering of words. The input to the algorithm is a text, which is a sequence of word $W_1, \ldots, W_n$. The output from the clustering algorithm is a binary tree, in which the leaves of the tree are the words. It runs time $O(V*K^2)$, where V is the size of the vocabulary and K is the number of clusters.

The Word Vector Class (WVC) model is induced via clustering the word vectors using a clustering algorithm. In order to build WVC model, we used word2vec to train word vectors of each word. Word2vec is a tool implemented by Mikolov et al. [19], which contains two distinct models, namely the CBOW which predicts the current word based on the context, and the Skip-gram which predicts surrounding words given the current word. Each model has two different training methods (with/without negative sampling) and other variations (e.g. hierarchical softmax), which amount to small "space" of algorithms. About the choice of the Training Parameters, [20] shows that in the case of a smaller corpus, a simpler model, such as Skip-gram, can achieve better results, whereas for a larger corpus, more complex models, such as CBOW and Order, are typically superior. And for the semantic property tasks, larger dimensions will lead to better performance. However, for the NLP tasks, a dimensionality of 50 is typically sufficient. Therefore, in order to generate a good word embedding, the training parameters are set as follows: 35 iterations, 250-dimension, window size is 5 and Skip-gram and hierarchical softmax are used. Then the word vectors are clustered using a K-means algorithm to drive a Word Vector Class (WVC) model.

4.3 Supervised learning

Conditional Random Fields (CRFs) is a probabilistic undirected graphical model which has been used successfully in many sequence tasks. Since CRFs have good description ability on the long-distance correlation that avoids the labeling paranoid problem, we apply the CRF tools to build the complex noun phrase recognition model.

5 Experiments

5.1 Dataset

For the word clustering features, we prepared 1.2 million Chinese sentences from

aircraft technical publications. Ten thousand Chinese sentences extracted from the aircraft technical publications are tokenized manually and then labeled according to the annotation specification in section 3 and then divided into two by the proportion of 8:2, one as training corpus, the other as testing corpus.

5.2 Evaluation measure

The metrics for evaluating complex noun phrase recognition models include precision rate, recall rate and their harmonic mean F score.

Precision measures the percentage of labeled complex noun phrases that are correct. Here "correct" means both the boundary of complex noun phrase and the label are correct. And the precision is therefore defined as:

$$\text{Precision} = \frac{\#\text{correct proposed tagged word}}{\#\text{correct complex noun phrase tags}} \tag{1}$$

Recall measures the percentage of complex noun phrase presented in the input sentences that are correctly identified by the system. Recall is defined as follows:

$$\text{Recall} = \frac{\#\text{correct proposed complex noun phrase tags}}{\#\text{current complex noun phrase tags}} \tag{2}$$

The F-measure illustrates a way to combine the previous two measures into one metric. The formula of F-score s is defined as:

$$\text{F} = \frac{\#(\beta^2+1)\times\text{Recall}\times\text{Precision}}{\#\beta^2\times\text{Recall}+\text{Precision}}, \beta = 1 \tag{3}$$

5.3 Experimental results

Besides the basic features, we incorporate the word clustering features trained by Brown clustering model and WVC model. The range of the number of word clustering is from 100 to 1700, and the step is 100. We started conducting a run with a basic feature setting, and gradually increased the complexity of the feature space for further runs. In general, clustering is an optimization procedure based on a specific clustering criterion, so clustering combination can be regarded as a technique that constructs and processes multiple clustering criteria rather than a single criterion [21]. So we assume that the combination of different word cluster features may work better than one word clustering feature. In order to verify our assumption, during the experiments, we not only use a single word clustering feature, but also the combination of different word cluster features. The experimental results are shown in table 5.

Table 5. Results of different runs with varied features

Features			Pre/%	Rec/%	F-scr/%
baseline			89.10	89.08	89.09
baseline +Brown 300			90.40	90.33	90.36
baseline +Brown 300	+Brown 700		90.65	90.93	90.79
baseline +Brown 300	+Brown 700	+Brown 1000	90.85	90.93	90.89
baseline +Brown 300	+Brown 700	+WVC 1500	90.79	91.03	**90.91**
baseline +Brown 300	+WVC 1100		90.60	91.07	90.84
baseline +Brown 300	+WVC 1100	+Brown 300	90.68	91.11	**90.90**
baseline +Brown 300	+WVC 1100	+WVC 1000	90.74	91.11	**90.92**
baseline + WVC 1600			89.95	90.06	90.00
baseline + WVC 1600	+Brown 200		90.58	91.01	90.79
baseline + WVC 1600	+Brown 200	+Brown 800	90.61	91.15	**90.88**
baseline + WVC 1600	+Brown 200	+WVC 500	90.50	91.13	**90.81**
baseline + WVC 1600	+ WVC 900		90.27	90.70	90.48
baseline + WVC 1600	+ WVC 900	+Brown 1100	90.53	90.93	**90.73**
baseline + WVC 1600	+ WVC 900	+WVC 1000	90.41	90.84	90.63

Feature groups are separated by (+). The parameters following Brown and WVC are the number of classes induced in each model. Pre: Precision; Rec: Recall; F-scr: F-score

1. When adding the single word clustering feature, we found that the performances of Brown model and the WVC model increased by 1.27% and 0.91% F-measure than baseline, respectively. This shows that besides the basic features, adding the word clustering feature can improve the recognition performance.

2. Besides the basic features and one word clustering feature, we add a word clustering feature again. For example, for the baseline +Brown 300, we add the Brown 700 and WVC 1100, respectively. The results show that the system with the baseline +Brown 300+Brown 700 and the baseline +Brown 300+WVC 1100 performed higher than the baseline +Brown 300 by 0.43% and 0.48%, respectively. Obviously, the combination of two word clustering features performed better than the combination of one word clustering feature. In addition, the combination of the Brown clustering model and the WVC model performed better than the combination of the Brown clustering model or the WVC model. The results also verify that our assumption is correct.

3. Based on the features above, we continue to add a word clustering feature, namely the combination of the basic features and three word clustering features. For example, for baseline +Brown 300+Brown 700, we add the Brown 1000 and WVC 1500, respectively. Results show that the system with the baseline +Brown 300+Brown 700+ Brown 1000 and the baseline +Brown 300+Brown 700+WVC 1500 performed higher than the baseline +Brown 300+Brown 700 by 0.1% and 0.12%, respectively. Obviously, the combination of three word clustering features tended to obtain higher F-score than the combination of two word clustering features. Although the increment is slight, the results are consistent with our assumption. Among the features, the combination of basic features and three word clustering features, namely the base-

line+Brown 300+WVC 1100+WVC 1000, achieved 90.92% F-score, which is 1.83% higher than the baseline.

6 Error analysis

Carrying on analysis on the complex noun phrases wrongly identified by the system, we observed that the following four types have a large proportion of all wrong examples.

1. A sentence containing “与” or “和” has accounted for about 8% of all wrong examples.

   Wrong result： CNP[侧/NN 撑杆/JJ 上端/NN 与/C 机身/NN] 相连/VV
   Correct result：CNP[侧/NN 撑杆/NN 上端/ NN] 与/P 机身/NN 相连/VV
   English: CNP[The upper end of the side] stay connects to fuselage

2. The left border or the first word before the left border is verb, accounting for 17.7% of all wrong complex noun phrase identified by the system.

   Wrong result: CNP[打开/VV 厨房/NN 区域/NN 天花板/NN]
   Correct result: 打开/VV CNP[厨房/NN 区域/NN 天花板/NN]
   English: open CNP[the ceiling in the galley area]

3. The right border or the first word after the right border is verb, accounting for 16.3% of all wrong complex noun phrases identified by the system.

   Wrong result: … 导致/VV CNP[控制面/NN 非指令/NN] 打开/VV
   Correct result: … 导致/VV CNP[控制面/NN 非指令/NN 打开/VV]
   English: … result in CNP[an inadvertent deployment]

4. Some gerunds are not labeled, accounting for 10.3% of all wrong complex noun phrases identified by the system.

   Wrong result: 在运输活体过程中，对于/P 过热/NN 保护/NN, …
   Correct result: 在运输活体过程中，对于/P [过热/NN 保护/NN], …
   English: during animal transportation for CNP[overheat protection] , …

From the above analysis, the solutions for the problems are: 1) improve the accuracy of the POS tagging, for instance revising the POS manually; 2) mine the semantic knowledge by other ways.

7 Conclusion and future work

Due to the large amount of complex noun phrases in the technical publications, we propose a method for complex noun phrase recognition. Through analysis on

the complex noun phrases in Chinese-English bilingual corpus from aircraft technical publications, we present an annotation specification to label the complex noun phrases in Chinese sentences based on the existing annotation specification. However, the annotation rules summarized in section 3 are not complete because of the limitation of the scale of corpus, some cases may not occur in our corpus.

For the complex phrase identification, we incorporate the word clustering features into the machine learning besides the word and POS features. Experimental results show that the combination of two word clustering features tended to achieve higher F-score than the combination of one word clustering feature, and the combination of three word clustering features tended to achieve higher F-score than the combination of two word clustering features. In addition, the combination of word clustering features trained by the WVC model and word clustering trained by the Brown clustering model performed better than the combination of word clustering trained by the WVC model or the Brown model. Among the features, the system with the combination of three word clustering features (the baseline +Brown 300+WVC 1100+WVC 1000), achieved 90.92% F-score, which is 1.83% higher than the baseline.

In the future, we would like to continue to explore other methods for feature combination and try to find new features for the complex noun phrase identification.

A Appendix

Table 6. Feature selection

Type	Feature
Atomic templates	W_{-4}, W_{-3}, W_{-2}, W_{-1}, W_0, W_1, W_2, W_3, W_4, P_{-3}, P_{-2}, P_{-1}, P_0, P_1, P_2, P_3
Composite templates	$W_{-4}W_{-3}$, $W_{-3}W_{-2}$, $W_{-2}W_{-1}$, $W_{-1}W_0$, W_0W_1, W_1W_2, W_2W_3, W_3W_4, $W_0W_{-1}W_{-2}$, $W_{-1}W_0W_2$, $W_0W_1W_2$, $P_{-3}P_{-2}$, $P_{-2}P_{-1}$, $P_{-1}P_0$, P_0P_1, P_1P_2, P_2P_3, $P_0P_{-1}P_{-2}$, $P_{-1}P_0P_2$, $P_0P_1P_2$, $W_{-1}P_{-1}$, W_0P_{-1}, W_0P_0, W_0P_1, W_1P_1, X, Y, W_{-4}X, W_4X, XY, B_0^{300}, B_0^{1600}, B_1^{300}, B_1^{1600}, B_{-1}^{300}, B_{-1}^{1600}, W_0 B_0^{300}, $W_{-1}B_{-1}^{300}$, $W_1B_1^{300}$, $W_0B_0^{1600}$, $W_{-1}B_{-1}^{1600}$, $W_1B_1^{1600}$, $W_0B_0^{300}B_0^{1600}$, W_{-1} $B_{-1}^{300}B_{-1}^{1600}$, $W_1B_1^{300}B_1^{1600}$, $P_0B_0^{300}$, $P_{-1}B_{-1}^{300}$, $P_1B_1^{300}$, $P_0B_0^{1600}$, $P_{-1}B_{-1}^{1600}$, $P_1B_1^{1600}$, $P_0B_0^{300}B_0^{1600}$, $P_{-1}B_{-1}^{300}B_{-1}^{1600}$, $P_1B_1^{300}B_1^{1600}$

W: word; P: pos; W_i: the current word; B^{300}: 300 Brown clustering, B^{1600}: 1600 Brown clustering; X: the fifth word before W_i; Y: the sixth word before W_i; W_{i-1}: the first word before W_i; W_{i+1}: the first word after W_i; B_i^{300}: 300 Brown clustering of W_i; B_i^{1600}: 1600 Brown clustering of W_i, and so on.

References

1. Xu Haifeng.: Application of Commercial Aircraft Technical Publication Specifications. J.

Aviation Maintenance & Engineering. 6, 91-93 (2012)
2. Zhou Qiang.: Annotation Scheme for Chinese Treebank. J. Journal of Chinese information. 18(4), 1-8 (2004)
3. T. Koo, X. Carreras, M. Collins: Simple semi-supervised dependency parsing. C. Proceedings of 46th Annual Meetings of the Association for Computational Linguistics (ACL). 595–603(2008)
4. Candito M, Crabbé B. Improving generative statistical parsing with semi-supervised word clustering. C. Proceedings of the 11th International Conference on Parsing Technologies. Association for Computational Linguistics, 138-141(2009)
5. Liang P. Semi-supervised learning for natural language. D. Massachusetts Institute of Technology (2005)
6. Zhu X, Goldberg A B. Introduction to semi-supervised learning. J. Synthesis lectures on artificial intelligence and machine learning, 3(1): 1-130,(2009)
7. Brown PF, deSouza PV, Mercer RL, Pietra VJD, Lai JC.: Class-Based n-gram Models of Natural Language.Compututational Linguistics. 18, 467-497(1992)
8. Lafferty J, McCallum A, Pereira F C N.Conditional random fields: probabilistic models for segmenting and labeling sequence data. J. 139-141 (2001)
9. Sun Ruina, Liu Qian. Chinese base noun phrase identification based on mutual information. J.China Computer & Communication. 11 (2012)
10. Wang Meng, Zhu Hong, Xu Yi. A study of automatic acquisition of Chinese compound noun phrases based on corpus. J. Journal of Leshan Teachers. 12 (2014)
11. Li Guochen, Dang Jianbing et al. Chinese base-chunk identification based on distributed character representation. J. Journal of Chinese information. 28(6), 18-25 (2014)
12. Zhang Kaixu, Zhou Changle.: Unsupervised feature learning for Chinese lexicon based on auto-encoder. J. Journal of Chinese information. 27(5):1-7 (2013)
13. Munkhdalai T, Li M, Batsuren K, et al.: Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. J. Journal of Cheminformatics. (2015)
14. Y.-C. Wu, A top-down information theoretic word clustering algorithm for phrase recognition. J. Information Sciences. 275, 213–225 (2014)
15. Zhu L, Chao L S, Wong D F, et al.: A noun-phrase chunking model based on SBCB ensemble learning algorithm. C. Machine Learning and Cybernetics (ICMLC), 2012 International Conference on. IEEE.11-16 (2012)
16. Konkol M, Brychcín T, Konopík M.: Latent semantics in Named Entity Recognition. J. Expert Systems with Applications. 42, 3470–3479 (2015)
17. Yu Shiwen, Duan Huiming, Zhu Xuefeng. The basic processing of contemporary Chinese corpus at Peking University. J. Journal of Chinese information Processing, 16(5): 49-64 (2002)
18. Wang Zhanqi. A contrastive study between English and Chinese of attributive-centered structure. D. Liaoning Normal University(2012)
19. Mikolov T, Chen K, Corrado G, Dean J: Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR (2013)
20. Lai, S., Liu, k., Xu, L., Zhao: How to Generate a Good Word Embedding?.J. arXiv preprint arXiv:1507.05523 (2015)
21. Qian Y, Suen C Y: Clustering combination method. C. 15th International Conference on IEEE, 2: 732-735 (2000)