Semantic Dependency Labeling of Chinese Noun Phrases Based on Semantic Lexicon

Yimeng Li[100083], and Yanqiu Shao[1*][100083], and Hongkai Yang[100083]

Beijing Language and Culture University, Beijing, China
liyimengblcu@126.com
yqshao163@163.com

Abstract: We have presented a simple algorithm to noun phrases interpretation based on hand-crafted knowledge-base containing detailed semantic information. The main idea is to define a set of relations that can hold between the words and use a semantic lexicon including semantic classifications and collocation features to automatically assign relations to noun phrases. We divide the NPs into two kinds of types: NPs with one verb or non-consecutive verbs and NPs with consecutive verbs, and design two different labeling methods according to their syntactic and semantic features. For the first kind of NPs we report high precision, recall and F-score on a dataset with nine semantic relations, and for the second type the results are also promising on a dataset with four relations. We create a valuable manually-annotated resource for noun phrases interpretation, which we make publicly available with the hope to inspire further research in noun phrases interpretation.

Keywords: Noun relations, Semantic dependency, Noun phrases.

1 Introduction

The automatic semantic dependency analysis of complex noun phrases such as "饮料/n 质量/n 监督/v 抽查/v" (the supervision and spot check of beverage quality), "文化/n 展示/v 活动/v 开幕式/n" (the opening ceremony of cultural exhibition) and "商业/n 调查/v 统计/v 工作/v" (business investigation and statistical work) is a difficulty in different languages. However, an important step towards being able to ascertain sentence meaning is to analyze the meaning of such NPs more generally. This kind of NPs have three basic properties which pose difficulties for their interpretation: (1) the compounding process is extremely productive; (2) the semantic relationship between the head and its modifier is implicit; (3) the interpretation can be influenced by variety of contextual and pragmatic factors[1]. The problem becomes more complicated when there are more than one "verbs" in the nominalization such as "节能/v 监督/v 管理/v 工作/v" (supervision and management of energy-saving) which is a noun phrase but constituted of "verbs", because verbs and verb phrases in Chinese can be nominalized without overt marker for it.

Semantic Dependency Analysis is a kind of deep semantic analysis, which describes the relations that hold between each word in a sentence. Therefore, basically the semantic dependency analysis of NPs is to explore the semantic relations that hold between nouns or verbal nouns. Interpretation of noun compounds (NCs) is highly dependent on lexical information[2]. One of the theme of SemEval-2007 is the interpretation of nominalizations and the researches based on different methods prove the importance of lexical resources[3]. So we explore the possibility of using an existing hand-crafted semantic knowledge-base (SKCC, Semantic Knowledge-base of Contemporary Chinese) for the purpose of placing words from a noun phrase into categories, and using the semantic classifications and collocation features to determine the relations that hold between nouns. The introduction of the semantic knowledge-base is aimed to add knowledge to the interpretation of complex NPs, because the process is highly dependent on encyclopedic knowledge. Our goal is to extract propositional information from the NPs, so the NPs extracted for the present study consist of at least one verbal noun.

[1] * Corresponding Author

In the following sections we will discuss the characteristics of the semantic knowledge-base SKCC that is used for the classification of semantic relations, the method to determine the relations, the evaluation of the results and some conclusions.

2 Related Work

We focus on the methods that analyze semantic relations in noun phrases. Rosario and Hearst[4] use a machine learning algorithm and a domain-specific lexical hierarchy achieving more than 60% accuracy on a dataset with 16 semantic relations. Hearst and Fillmore[2] continue the study by placing the words from a two-word compound from biomedical domain into categories, and then using this category membership to determine the relations that hold between nouns, obtaining classification accuracy of approximately 90% on a dataset with 35 semantic relations. Girju and Moldovan[5] use machine learning tools such as SVM, semantic scattering to explore noun relations with attributes extracted from ComLex and VerbLex . It turns out that SVM achieves the highest accuracy of 72% on a dataset with 35 semantic relations. Nastase and Szpakowicz[6] use the machine learning tools such as decision trees, instance-based learning and Support Vector Machines, with attributes respectively extracted from WordNet and corpus. The F-measure reaches a maximum of 82.47% on a dataset with 5 semantic relations after the introduction of the WordNet word meaning. Tratz and Hovy[7] use Maximum Entropy Classifier and $SVM^{multiclass}$ to classify semantic relations between English noun compounds and obtain an accuracy of 79.3% and 79.4% respectively. Their dataset comprises 17509 compounds with a new taxonomy of 43 semantic relations. Dima and Hinrichs[8] use neutral network classifier implemented in the Torch7 scientific computing framework for the automatic classification of noun compound semantic relations. The F-measure reaches a maximum of 79.48% on a dataset with 12 semantic relations.

Lijie Wang[9] focuses on the automatic analysis of Chinese sentences using Graph-based algorithm combined with knowledge from dependency parsing. The experiment achieves an accuracy of 66.83% on a dataset with 19 semantic relations. Ding Yu[10] automatically labels Chinese sentences based on the dependency graph. On the basis of the results obtained from dependency tree, she uses SVM to construct the dependency graph and achieves an accuracy of 69.37% on a dataset with 32 relations.

3 The Semantic Classification: SKCC

The semantic interpretation of complex NPs requires a great deal of world knowledge, and semantic lexicon can provide a certain background knowledge for automatic semantic analysis. There are mainly two forms of semantic knowledge[11]: (1) category knowledge, which can be expressed in the form of "attribute: value". For example, the simple description of the semantic attribute of a word: [semantic class: food] or the description of the semantic relations between two or more words: {Agent [semantic class: people | animal]}; (2) rule- based knowledge, which can be described in the form of "condition -> action". If the semantic properties of a noun and the semantic requirements of a verb's object argument are consistent, then the noun and the verb can make a verb-argument structure. The above category knowledge, together with rule-based knowledge, can be used to explain the difference between "eating apples" and "eating circles" (unacceptable). The dictionary-based approach is designed to use the semantic category to determine the acceptability of phrases or sentences.

The Semantic Knowledge-base of Contemporary Chinese (SKCC) is a large scale of bilingual semantic resource developed by Chinese Department of Peking University. It provides quite amount of semantic information such as semantic classification and collocation features for 66539 Chinese words and their English counterparts[12, 13]. It has three dictionaries including noun, adjective and verb. Each dictionary is organized in hierarchy with respect to semantic classes. Nouns are divided into four general semantic classes: entity, abstraction, time and space, and each of the category is further divided into multiple subcategories. The noun dictionary has three components: the word, the part of speech and the semantic class. For example: "金属" (metal), (metal, n, material). The verb dictionary is a six-point group: the word, the part of speech, the semantic class, the argument quantity, the subject and the object. For example the verb "切削" (cut), (cut, v, body function, 2, person, entity). Here is an example for analyzing, "金属/n 切削/v 机床/n" (metal cutting machine tool). Firstly we respectively map the words of the NP into the noun and verb dictionary, and then match the semantic class of "金

属” (metal) to that of the subject and object of “切削” (cut). It turns out that “金属” (metal) could be an object of “切削” (cut), because the semantic class of “金属” (metal), i.e., “material” is a hyponym of the object of “切削” (cut) , i.e., “entity”. The specific semantic label will be further determined according to the semantic classes of the nouns and verbs. The method will be elaborated in section 4.

4 Method and Evaluation

4.1 The Dataset of NPs

In the paper, we divide the noun phrases into two types: NPs with one verb (verbal noun) or non-consecutive verbs and NPs with consecutive verbs. The dataset for the study of noun phrases with one verb or non-consecutive verbs consists of 1035 noun phrases, and the one for NPs with consecutive verbs consists of 525 noun phrases. These NPs are automatically extracted from BLCU-HIT Semantic Dependency Graph Bank and Dynamic Circulation Corpus (DCC) and then checked manually. We run a part-of-speech tagger on LTP (Language Technology Platform) and a program that extracts only sequences of units tagged as nouns and verbs. The data are further inspected manually to create a dataset of only NPs.

4.2 Labeling NPs with one “Verb” of Non-consecutive “Verbs”

Firstly we will discuss the annotation method of NPs with one “verb” or non-consecutive “verbs” such as “尾气/n排放/v标准/n” (vehicle emission standard) and “垃圾/n运输/v车辆/n管理/n” (waste vehicles management). Given a NP with one “verb” or non-consecutive “verbs”:

a. map the nouns and verbs to the noun and verb dictionary respectively;
b. match the verbs and arguments based on collocation features;
c. choose specific semantic labels according to the semantic classes of nouns and verbs;

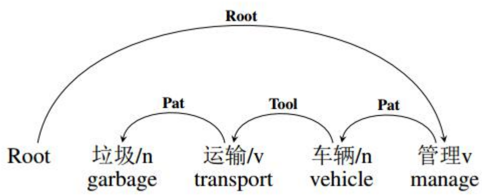

Fig. 1. An example for NPs with one verb or non-consecutive verbs.

As is illustrated in Fig.1, first we map the nouns “垃圾” (garbage) and “车辆” (vehicle) to the noun dictionary finding their semantic classes respectively fall into “artifact” and “tool” ; and then map the verbs “运输” (transport) and “管理” (manage) to the verb dictionary getting following information:
(transport, v, motion, 2, person | vehicle, entity),
(manage, v, other event, 2, person, entity | abstraction | process)

Second we identify that “垃圾” (garbage) and “车辆” (vehicle) are the objects of the verbs “运输” (transport) and “管理” (manage) respectively, because the semantic classes of “垃圾” (garbage) is “artifact” which is a subclass of the object of “运输” (transport), i.e., “entity”, and similarly the semantic class of “车辆” (vehicle) is “tool” which is a subclass of the object of “管理” (manage), i.e., “entity”. Lastly we identify the specific labels according to the semantic classes of words, as is shown in Table 1. We think that the relations between words can be tentatively displayed by the word meaning. Take the subject as an example, organism that can act consciously could be an “Agent” while the subject that is described is an “Experiencer”. Similarly, verbs in the SKCC are divided into 13 categories according to the semantic classes, and we design different relations such as “Patient”, “Content” and “Product” according to the meaning of the verb. For example, “部门交流” (communication between departments) and “部门直属” (directly subordinate to), since the semantic class of the subject are the same, we turn to that of verbs. “交流” (communication) falls into “Communication” while “直属” (directly subordinate to) falls into “State”. It is true that “Communication” has to be done intentionally while “State” is a description of state, and thus the first relation is “Agt” and the second is “Exp”. Therefore, The specific semantic label is determined by the semantic classes of nouns and verbs of the NPs.

Table 1. Semantic labels categorized by semantic classes.

Semantic Class of Nouns				Rel	Remark
1Things	1.1Entity	1.1.1 Organism	1.1.1.1 Person	Agt/Exp/Aft, Pat/Cont/Link	
			1.1.1.2 Animal	Agt/Exp/Aft, Pat/Cont/Link	
			1.1.1.3 Plant	Exp, Pat/Cont/Prod/Link	
			1.1.1.4 Microbe	Exp, Pat/Cont/Prod/Link	
		1.1.2 Object		Exp, Pat/Cont/Prod/Link	Tool： Tool
		1.1.3 Part		Exp, Pat/Cont/Prod/Link	
	1.2Abstraction	1.2.1Attribute 1.2.2Info 1.2.3Field ……		Exp, Pat/Cont/Prod/Link	
2Time				Time	
3Space				Loc	

4.3 Labeling NPs with Consecutive "Verbs"

The NPs with consecutive "verbs" in Chinese such as "学生/n 申诉/v 处理/v 制度/n" (processing system of student's appeal) , "疾病/n 预防/v 控制/v 中心/n" (centers for disease control and prevention) and "分管/v 工作/v 完成/v 情况/n" (the completion of the work that is in charge of) is a common linguistic phenomenon. These are nominalizations with verbal nouns. Because there is no inflection in Chinese, the part-of speech taggers of the verbal nouns are also "verbs", which poses a difficulty for automatic acquisition of semantic relations. After annotating 525 such NPs manually, we find that there are four kinds of relations that hold between the verbal nouns.

a. Case Relation: one of the "verb" is an object of the other, such as "污染/v 处理/v 设备/n" (anti-pollution devices), "电子/n 产品/n 污染/v 控制/v 管理/v 办法/n" (measures of the management and control of the electronic product), "防汛/v 抢险/v 物资/n 运输/v 保障/v 演习/v" (the exercise of ensuring the transportation of relief supplies), as is illustrated in Fig. 2. The semantic labels are "Pat", "Cont", "Prod";
b. Attribute Modifier: one of the nominalization is a modifier of the other, such as "加工/v 贸易/v 比重/v" (portion of processing trade), "上网/v 服务/v 营业/v 场所/n" (business site for Internet service), "节水/v 教育/v 读本/n" (water-saving education reader), as is illustrated in Fig. 3. The semantic label between them is "Desc";

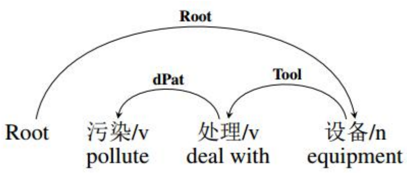

Fig. 2. Case Relation.

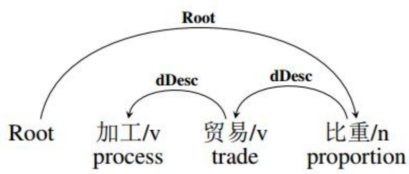

Fig. 3. Attribute Relation.

c. Coordinating Relation: the two nominalizations function equally syntactically, such as "疾病/n预防/v控制/v中心/n" (centers for disease control and prevention), "投诉/v举报/v中心/v" (complaint center), "生产/v加工/v企业/n" (production and processing enterprises), as is illustrated in Fig. 4. The semantic label between them is "eCoo";
d. Adverbial Modifier: the former nominalization describe the manner when the latter act is done, such as "全封闭式/n 无菌/v 操作/v" (totally enclosed aseptic operation), "燃煤/v 采暖/v 锅炉房

/n ” (coal heating boiler), “跨境/v 采访/v 报道/v 活动/v” (cross border coverage), as is illustrated in Fig. 5. The semantic relation between them is “Mann”.

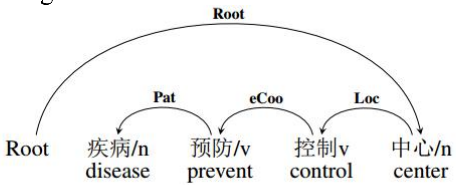

Fig. 4. Coordinating Relation.

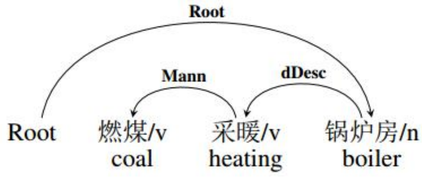

Fig. 5. Adverbial Relation.

The method to label the NPs with consecutive “verbs” is a little different from the one mentioned above as is shown in Fig. 6:

a. map the nouns and verbs to the noun and verb dictionary respectively to obtain the semantic classes;
b. match the semantic class of the object of each “verb” with that of the noun before them;
c. if the noun could be an argument of all the “verbs”, the relation that holds between them is labeled “eCoo” ;
d. if the semantic class of the noun’s and that of the “verbs” can not match totally, then eliminate the last “verb” and continue to match the semantic class of the noun with that of the rest of objects until all the objects are considered;
e. if there is no case relation found after all the objects are considered, the first “verb” is degraded as an “abstraction” noun, and then identified if it could be an argument of the latter “verbs”; this process is like the one in step b;

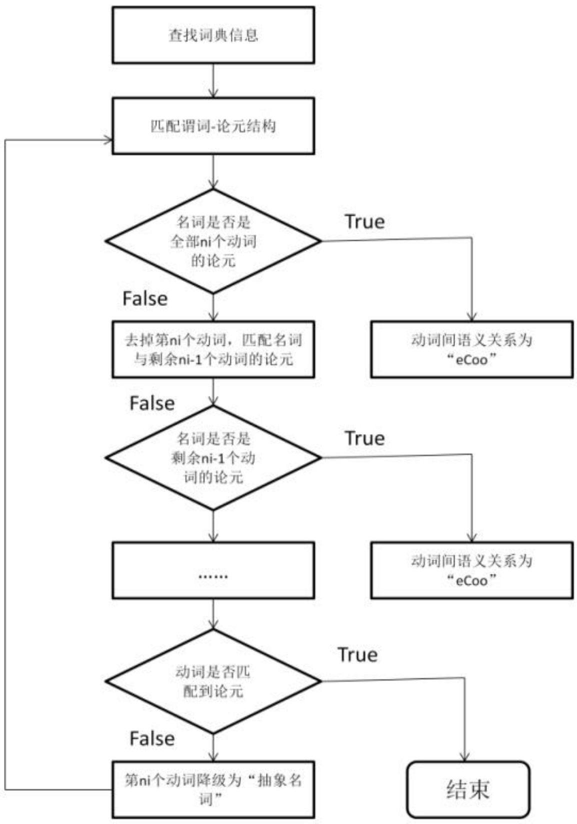

Fig. 6. The method to label the NPs with consecutive “verbs”.

For example, “电子/n 产品/n 污染/v 控制/v 管理/v 办法/n” (measures of the management and control of the electronic product), firstly, we map the nouns and verbs of the NPs into dictionaries to obtain their semantic classes. Here are the semantic classes these words fall into:

(电子 (electronics), n, substance)
(产品 (product), n, artifact)
(污染"(pollute), v, other event, 2, entity, space | natural object)
(控制 (control),v, other event, 2, person, entity | abstraction | event)
(管理 (manage), v, other event, 2, person, entity | abstraction | process)
(方法 (measure), n, process)

Secondly, we match the semantic class of the noun "产品" (product) with that of the objects of the three "verbs", and it turns out that "产品" (product) is an subject of "污染" (pollute), while it is an object of other "verbs". Since the relations between the noun and the "verbs" are not the same, the first "verb" "污染" (pollute) is degraded as an "abstraction" noun which could be an argument of both "控制" (control) and "管理"(manage) after being matched with the semantic classes of the rest of objects, as is shown in Fig. 7.

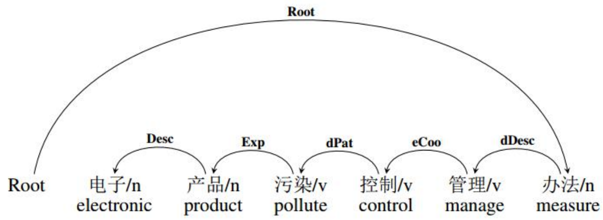

Fig.7. An example of NPs with consecutive "verbs".

It is worth noting that the dependency arcs here are set to "Division" [14] by default, which means that the current word's father node is the word after it. And then the dependency arcs will be adjusted according to the labeling of relations.

4.4 Accuracies of NPs with One "Verb" or Non-consecutive "Verbs"

We automatically extracted 1035 NPs with one verb or non-consecutive verbs, and use the algorithm mentioned above to automatically predict which relations should be assigned. We checked the results manually and found high accuracies in most of the relations, as is shown in Table 2.

Table 2. The accuracy, recall and F-measure of the NPs with one "verb" or non-consecutive "verbs".

Rel	Occur	Accuracy	Recall	F-measure	Examples
Agt	214 (16.7%)	79.18	81.77	80.45	管理部门
Pat	479 (37.6%)	90.75	65.55	76.11	垃圾处理
Exp	123 (9.6%)	93.33	68.29	80.37	经济运行
Prod	104 (8.2%)	90.47	91.34	90.89	汽车生产
Cont	72 (5.5%)	87.23	56.94	68.90	天体研究
Mann	84 (6.5%)	64.70	39.28	48.41	规模经营
Tool	21 (1.5%)	65.00	61.90	63.41	切削机床
Loc	161 (13%)	95.13	85.09	89.83	北京总部
Time	17 (1.3%)	80.00	70.58	74.99	雪天运输

Because there is no similar research in Chinese, the baselines we choose are from Lijie Wang[8] and Yu Ding[9] who analyze the whole Chinese sentence using Graph-based algorithm and SVM. We select five intersection tags (three tags of Graph-based method) to compare the results as is shown in Fig. 8. The task of semantic labeling the whole sentence is absolutely more difficult because the components, syntax structure and semantics of sentences are more complicated while the input in our program are only noun phrases through manual inspection which means it is easier to analyze, so our method based on rules performs better.

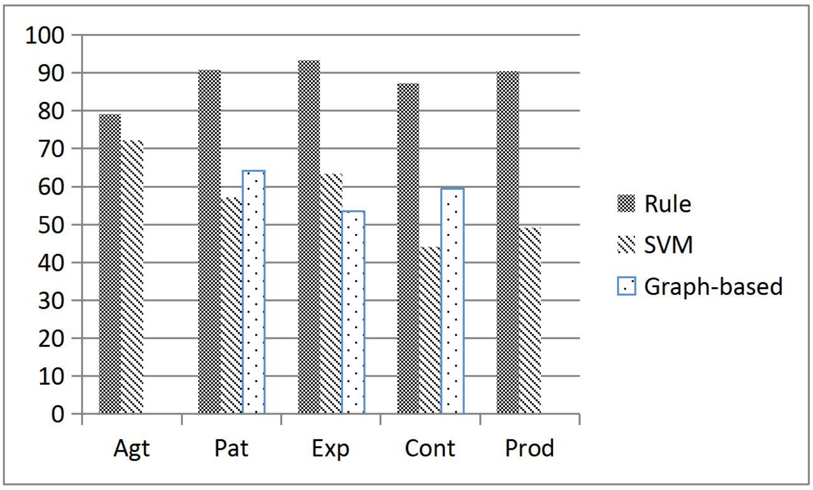

Fig. 8. The accuracy, recall and F-measure of the methods based on rules, SVM and Graph-based algorithm.

Generally the method based on semantic lexicon performs well when it comes to the identification of case relation, because in our case this is done literally according to the semantic class and collocation features. There are three main factors affecting the accuracy of the case relation. Firstly, the labeling of this relation is highly dependent on the properties of the lexical hierarchy and collocation information. The case relation can not be found if the two words' collocation information does not match. For example, "废水收集装置" (waste water collection device), the semantic class of "废水" (waste water) falls into "attribute" while that of the object of "收集" (collect) is "entity"; Secondly, the present program is not able to deal with the reverse relations which means the head argument lies behind the verb, such as "接待/v 游客/n 数量/n" (the number of tourists); Thirdly, it is hard to identify the case relation between words without direct semantic arcs. For example, "土地/n 规模/n经营/v" (land-scale management) , "土地" (land) is an patient of "经营" (manage), but the program fails to identify because the semantic arcs are set "Division" by default.

Time and Loc also have high accuracies because the labeling of these are highly dependent on word meaning. As for the labeling of relation Tool and Mann, the accuracy is pretty low, so it proves that semantic lexicon helps little in identifying these relations of which interpretation highly depends on contexts and pragmatics.

4.5 Accuracies of NPs with Consecutive Verbs

As for the NPs with consecutive "verbs", in order to label such kind of NPs we extracted 525 such NPs. Following are the results checked by hand as is shown in Table 2. The distribution of the four relations is equal overall. The relation "Desc" has the largest proportion accounting for 32.5%. It is an coincidence that Case Relation and Coordinating Relation have the same number due to the data scarcity. The Adverbial Relation accounts for a minimum proportion of 14.9%.

Table 3. The accuracy, recall and F-measure of the NPs with consecutive verbs.

Rel	Occur	Accuracy	Recall	F-measure	Examples
Case Rel	148 (26.3%)	72.88	58.10	64.65	支付保障
Desc	182 (32.5%)	76.73	54.96	64.04	节水教育
Coo	148 (26.3%)	68.70	68.24	68.46	生产经营
Mann	84 (14.9%)	60.75	55.95	58.25	应急处置

The result is pretty promising considering the difficulty of the task. The use of SKCC plays a significant role in the differentiation of relations between nominalizations. The factors affecting the Case Relation are basically the same with the NPs with one "verb" or non-consecutive "verbs". The main factor affecting the accuracy of Coordinating Relation firstly is the properties of collocation information. For example, "固定/a 资产/n 监督/v 管理/v" (the supervision and management of fixed assets) the relation between "监督" (supervision) and "管理" (management) is coordination, but the "资产" (assets) and "监督" (supervision) has no case relation according to the collocation information. So the program fails to identify the correct relation between "监督" (supervision) and "管理" (management). Besides, some times words with semantic classes being different could be coordinating. For example, "宣传/v 教育/v 活动/v", though the objects' semantic classes of "宣传" (advocate) and "教育" (educate) fall into different classes, the two words are coordinating syntactically and semantically . As for Adverbial Relation we simply label the verbs that fall into semantic classes (Other Event, 1, Person) and (Change, 1, Person) "Mann", so the accuracy is not very ideal.

5 Conclusion

We have presented a simple algorithm to noun phrases interpretation based on semantic lexicon. The main idea is to define a set of relations that hold between the words and use a semantic lexicon with semantic classification and collocation features to automatically assign relations within noun phrases. We divide the NPs into two kinds of types, and respectively design annotation methods for them according to their structure features. Through annotating the NPs manually we find that there are four kinds of relations between Chinese verbal nouns: Case Relation, Coordinating Relation, Attribute Relation and Adverbial Relation and we further propose a method based on the semantic lexicon to automatically assign relations for such NPs. The method performs well when it comes to the identification of case relations but the performance is not very ideal as for the identification of "Mann" and "Tool" of which interpretation more depend on contexts and pragmatics. The semantic labels for the present study is not sufficient enough due to the limit of method, however our purpose is to create a manually annotated dataset of Chinese complex NPs which is used to provide support for machine learning. So our next job will be to explore the semantic relations of complex NPs through machine learning. We hope the combination of rules and machine learning could move the complex NPs a step further towards being generally understood. Understanding relations between multiword expressions is important for many tasks, including question answering, textual entailment, machine translation and information retrieval among others.

Acknowledgments

Thanks to National Natural Science Foundation of China (NSFC) via Grant 61170144, Major Program of China's National Linguistics Work Committee during the twelfth five-year plan (ZDI125-41), Young and Middle Aged Academic Cadre Support Plan of Beijing Language and Culture University (501321303), Graduate Innovation Foundation in 2017 (17YCX137).

References

1. Lapata, M.: The Automatic interpretation of nominalizations. In: 17th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 716-721. AAAI Press (2000).
2. Rosario, B, Hearst, M., Fillmore, C.: The descent of hierarchy, and selection in relational semantics. In: Meeting on Association for Computational Linguistics, pp. 247-254. Association for Computational Linguistics, Philadelphia (2002).
3. Girju, Nakov,et, al.: SemEval-2007 Task 04: Classification of semantic relations between nominals. In: 4th Proceeding of International Workshop on Semantic Evaluations (SemEval-2007), pp. 13-18. Association for Computational Linguistics , Prague (2007).
4. Rosario, B, Hearst, M., Fillmore, C.: Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: Lee, L., Harman, D. (eds). In: Proceedings of EMNLP (Empirical Methods in Natural Language Processing), pp. 247-254. (2001).
5. Girju, R., Giuglea, A. M., Olteanu, M., et al.: Support vector machines applied to the classification of semantic relations in nominalized noun phrases. In: Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics. Association for Computational Linguistics, pp. 68—75 (2004).
6. Nastase, V., Sayyad-Shirabad, J., Sokolova, M., et al.: Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In: National Conference on Artificial Intelligence, pp. 781-786. AAAI Press (2006).

7. Tratz, Stephen, Hovy, et al.: A Taxonomy, dataset, and classifier for automatic noun compound interpretation. In: 48th Proceeding of the Annual Meeting of the Association for Computational Linguistics, pp. 678-687. (2010).
8. Dima, Hinrichs: Automatic noun compound interpretation using deep neutral networks and embeddings. In: 11th Proceeding of the International Conference on Computational Semantics, pp. 173-183. (2015).
9. Lijie, W.: Research on Chinese semantic dependency analysis. Doctoral dissertation. Harbin Institute of Technology. (2010).
10. Yu, D.: Dependency Graph based Chinese semantic parsing. Doctoral dissertation. Harbin Institute of Technology. (2014).
11. Weidong, Z.: Principles of determining semantic categories and the relativity of semantic categories. World Chinese Teaching (2), 3-13 (2001).
12. Hui, W.: Structure and Application of the semantic knowledge-base of modern Chinese. Applied Linguistics 2 (1), 134-141 (2006).
13. Hui, W, Weidong, Z.: New progress of the semantic knowledge-base of contemporary Chinese. In: 7th Joint Academic Conference on Computational Linguistics, Harbin (2003).
14. Li Y, Shao Y.: Annotating Chinese noun phrases based on semantic dependency graph. 21st International Conference on Asian Language Processing, pp. 18-21. IEEE, Tainan (2016).