*International Journal of Knowledge* *www.ijklp.org*
*and Language Processing* KLP International ©2012 ISSN 2191-2734
Volume 3, Number 1, **2012** pp.1—14

# Mining Commonsensical Semantic Relations from Noun-Noun Phrases

Shi Wang[1], Fei Xia[1], Yanan Cao[1], Yajun Pei[2], and Cungen Cao[1]
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Zhongguancun, Beijing 100080, China
wangshi@ict.ac.cn

[2] China National Committee for
Terms in Sciences and Technologies
Beijing, China
peiyj@cnctst.gov.cn

Received January 2012; revised January 2012

ABSTRACT. *Semantic knowledge is considered the bottleneck for many great challenges in artificial intelligence. Semantic relations, which are major components of knowledge, are traditionally mined from text using sentences-oriented lexico-syntactic patterns. However, most of commonsensical relations are not expressed in sentences at all because they are too familiar for people to be written formally. These commonsensical relations are often expressed in phrases. In this paper, a phrases-oriented approach is proposed to automatically mine semantic relations from noun-noun phrases. Firstly, a small group of seed phrases which contains specific relations are detected automatically with the aid of lexico-syntactic patterns. Then other phrases are judged whether contain same relations with the seeds or not by analogizing. Selected accepted phrases can be seen as new seeds and then more and more phrases can be judged. The analogizing process is based on hypothesis that similar phrases contain similar semantic relations. Four novel different kinds of similarities evaluating strategies, including symmetrical and asymmetrical WordNet-based ones, and symmetrical and asymmetrical SimRank-based ones, are put forward and compared in this paper. Experiment focusing on part-whole relation and Chinese noun- noun phrases shows that the method can achieve a precision of 0.91 and recall of 0.69.*

**Keywords:** sematic relations mining, phrases analysis, knowledge discovery

1. **Introduction.** Semantic knowledge is considered bottleneck for many heavily challenged human-level intelligences including natural language understanding, text mining, sentiment analysis, etc[1-3]. Semantic relations are major components of knowledge. Nowadays, the major approach to automatically mine semantic relations is

sentences-oriented lexico-syntactic patterns method [4-6]. However, there exist lots of commonsensical relations cannot be mined in this way because they do not appear in sentences at all. Most of such relations, which are too commonsensical for people to be written formally in sentences, are often concealed in phrases.

For example, we cannot even find one sentence from web by Google using query "paper cup is made of paper" or other similar queries, but there are more than millions of webpages for phrase “paper cup”. In this case, it is more efficient to mine commonsensical knowledge "PART-OF(paper, paper cup)" from phrases.

In this paper, we aim to mine large-scale commonsensical semantic relations from noun-noun phrases. Not considering all kinds of semantic relations, this paper only focuses on one specific relation $r$ such as IS-A, PART-OF, etc. Formally, we denote a noun-noun phrases as $t=w^1w^2$, where $w^1$ and $w^2$ are the two nouns of $t$. Given a set of phrases T={$t$}, function $f$ :T→{*true*, *false*} reflects whether phrases contain $r$ or not. We call $t$ a $r$-positive phrase if $f(t)$=*true*, or $r$-negative phrases otherwise. Given $t$, our goal is to calculate $f(t)$ and then mine semantic relations from it.

An absolutely hand-free approach was proposed for the task in this paper. In short, we anchored a small group of $r$-positive phrases (which were called seeds in our work) with the aid of lexico-syntactic patterns like "$w^1$ is a part ($t$|$w^2$)" firstly, and then, by analogizing word by word, gathered similar phrases and calculated their possibility to be $r$-positive ones. For instance, when affirmed "wood house is made of wood" by retrieving such sentences from web, we would tell "paper cup" or "plastic plate" contain PART-OF relations too for they were similar with "wood house".

Similarities evaluating between words was the crucial procedure. Four novel different kinds of similarities evaluating strategies, including symmetrical and asymmetrical WordNet-based ones, and symmetrical and asymmetrical SimRank-based ones, are put forward and compared in this paper.

The rest of the paper is organized as follows. Section 2 described related work. Section 3 gives the four different similarities calculating methods in detail, and section 4 presents phrases analogizing algorithm. Experiment result and discussion are given in section 5. Finally section 6 concludes and presents the future work.

2. **Related Works.** Although a considerable amount of work has been done on semantic relation detection, the dominant approach for the task is based on lexico-syntactic patterns. Hearst M. A. firstly developed the method for automatic acquisition of IS-A relations from sentences [4]. Lexico-syntactic patterns cannot be used for phrases because words are adjoined together without boundaries like "is a kind of", and boundaries are necessary in lexico-syntactic patterns.

For phrases semantic analyzing, [7] presents a semantic analysis method for Japanese $N_1$ の $N_2$ (roughly mean $N_2$ of $N_1$) noun phrases. The method is based on a decision tree classifier. They firstly conclude 36 kinds of semantic relations by hand, and then train a decision tree to analyze unseen phrases using a thesaurus and 19, 500 annotated phrases. Only one semantic relation is mined from a phrase in this work. This method achieves an accuracy of about 0.8. This supervised classifying method is widely used by many researchers, leaving aside the differences between semantic relation classes, the phrases similarities measurement, and classifiers.

[8] focuses on biomedical double-word noun phrases. They give 38 types of relations and adopt neural networks as classifier. Similarities between phrases are calculated based on a domain-specific lexical hierarchy. [9] presents 43 semantic relations in noun-noun phrases and annotates 17,509 phrases. They adopt WordNet-based features, Roget's thesaurus-based features, surface-level features, and N-gram features to train a SVM classifier. Multi semantic relations are extracted from one phrase. Their resources and results are available via http://www.isi.edu.

In [10], more than one relation is mined from one phrase. They focuses on 14 kinds of semantic relations and combine two methods, that is, dictionary based method and semantic feature based method, to analyze Japanese noun phrases. If multi different semantic relations are detected by the two methods, they accept all. Like their work, our approach can also detect multi relations from one phrase.

Ontologies with IS-A relations is very helpful for this problem. [8] adopts the juxtaposition of category membership within the lexical hierarchy when determining the relation that holds between pairs of nouns. They obtain 0.9 accuracy overall. Their work is limited to biomedical domain, so a comprehensive hierarchy is available.

In domain free applications, WordNet [11] is widely used. [12] presents an algorithm for automatically disambiguating noun-noun compounds by deducing the correct semantic relation between their constituent words. The algorithm takes as input the WordNet senses for the nouns in a compound, finds all part-whole relations of those senses, and searches the corpus for other compounds containing any pair of those senses. The relation with the highest proportional co-occurrence is returned as the correct relation for the compound.

Ideal lexicons or ontologies are very difficult to build. When such resources are inaccessible, linguistic features and corpus are used. [13] presents an approach for detecting semantic relations in noun phrases. They firstly identify and study the characteristics or feature vectors of each noun phrase linguistic pattern, then develop models for their semantic classification. [14] proposes four coarse-grained semantic roles of the noun modifier and use a Maximum Entropy Model to label such relations in a compound nominalization. The feature functions used for the model are corpus-based statistics acquired via role related paraphrase patterns, which are formed by a set of word instances of prepositions, support verbs, feature nouns and aspect markers.

3. **Calculating Similarities between Phrases.** [15] presents a cognitive research for semantic structure of noun phrases and puts forward an analogy hypothesis, that is, similar noun phrases have similar semantic structures. A group of similar noun phrases

such as "brick house", "plastic knife", "paper cup" and "silver bullet", have same semantic structure, i.e. MADE-OF($t$,$w^1$). The fundamental question under the hypothesis, which is also the key problem in our method and we have not touched so far, is under what circumstances, can we claim or disclaim two phrases are similar.

Calculating similarities between phrases can be implemented by two steps. The first step is calculating similarities between their corresponding words, and the second one is merging these similarities in some ways.

Similarities between words, which are ill defined, include two categories, i.e., taxonomic ones and associative ones [16]. The former is about cognitive taxonomy, whereas the latter focus on topics. Expected similarities in our task are commonly considered to be taxonomic similarities.

Despite the usefulness of taxonomic similarity measurements in many applications, a robust approach still remains a challenging task nowadays. Traditionally, there are two kinds of methods, i.e., thesauri-based and context-based methods, for the problem. Thesauri-based methods rely on existing thesauruses such as WordNet and HowNet, and context-based methods are founded on distributional hypothesis that "similar words appear in similar context". We designed two novel methods for comparison in this work. One utilized WordNet and the other take collocations as words' context.

Before we start presenting the two methods, we discuss symmetry of similarities. Being similar is commonly considered to be a symmetrical relationship. However, from the cognitive view of point, [17] provides empirical evidence to demonstrate that similarities should be asymmetric, and less salient concepts are more similar to the salient ones. For example, we prefer to say "leopard is like cat" rather than say "cat is like leopard", because cat has more salient stimulus for people and leopard has less ones. In the following two methods, we also considered symmetry of similarities.

3.1. **WordNet-based word similarities.** WordNet [11] is a large-scale, implicit, and widely used domain free ontology. It is a network in which nodes are concepts and edges present predefined relations such as PART-WHOLE, PART-OF, DOMAIN, ANTONYM, etc. A concept in WordNet, which is also called a synset, is a group of synonyms. WordNet is reconstructed in many other languages including Germany, French, Italian, Korea and Chinese [18-20].

WordNet helps a lot when calculating semantic similarities. Using it, many efforts have been done to calculating similarities between words, especially nouns. In cognitive science, [11] also proves that geometry-based cognitive model has the best cognitive result in a hierarchical concepts space. Accordingly, the dominant in these work, path-based approaches, claim that similarities between noun concepts can be reflected by the IS-A path length between them [21-24].

It is noticeable that in WordNet, a word with multi senses will be included in multi concepts. When word sense disambiguation is not handled, maximal similarities between concepts are often taken as similarities between these words. We follow this rule in this work.

### Symmetrical WordNet-based similarities

As a headmost work, [21] considers that similarities between concepts in WordNet are inversely proportional to the length of there is-A path. [22] further considers specificity of nodes. A concept is more specific if it is closer to leaves (or instances) in the taxonomy tree. [23] gives a nonlinear formula that performs nearly at a level of human replication and achieves a correlation of 0.901.

Beyond the above remarkable work based on IS-A path, we noticed that balances of concepts are also important. The balance of two concepts is their depth span in WordNet taxonomy tree. The reason to consider balance is we noticed that balanced words are more likely to form similar phrases. Fig.1 gave an example to explain this.

Based on an overall consideration of IS-A path distances, specific of concepts, and balance of concepts, we designed a novel symmetrical formula shown in Eq.(1).

$$sim_1(w^1,w^2) = \frac{0.5\times \mathrm{H(p)}}{D(w^1,w^2)+0.5\times \mathrm{H(p)}} \times \frac{1}{1+0.2\times \mathrm{B}(w^1,w^2)} \tag{1}$$

Where D($w^1$,$w^2$) is the IS-A edges number of the shortest path from concept containing $w^1$ to concept containing $w^2$. Concept $p$ is the most specific common hypernym concept of $w^1$ and $w^2$, and H($p$)=D($r$, $p$) is the depth of $p$ where $r$ is the root of WordNet taxonomy tree. B($w^1$,$w^2$)=|H($w^1$)-H($w^2$)| is balance degree of the two words.

In $sim_1$, two words in a same concept have a similarity of 1, and similarity between two brother concepts (concepts that have common direct hypernym) is H(p)/(4+H(p)).

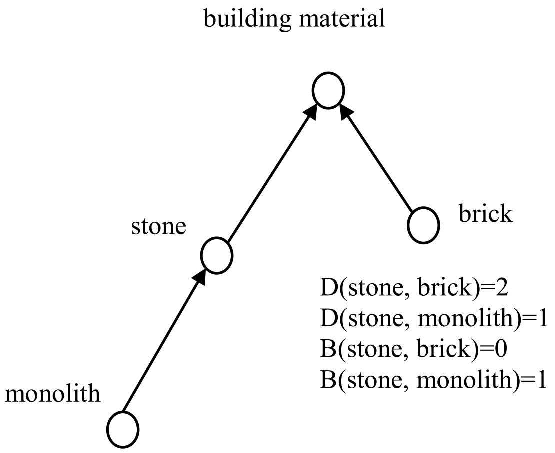

FIGURE 1. This is an example of balance influence in WordNet. "stone house" is more likely to contain similar relations with "brick house" than "monolith house", even though "monolith" is more near and more specific to "stone" than "brick".

### Asymmetrical WordNet-based similarities

According to [17], a less salient concept is more similar with a more salient one. In WordNet, we considered a concept is more salient if it has more neighbors. The basis of this hypothesis is that a more salient concept should get more attention, and then be related with more concepts. Based on this hypothesis, we developed another novel asymmetrical WordNet-based similarities measurement.

$$sim_2(w^1 \rightarrow w^2) = \frac{sim_1(w^1,w^2)}{1+{N(w^1)}/{N(w^2)}} = \frac{sim_1(w^1,w^2)\times N(w^2)}{N(w^1)+N(w^2)} \qquad (2)$$

Where $sim_2(w^1 \rightarrow w^2)$ means the degree that how $w^1$ is similar with $w^2$. N($w^1$) and N($w^2$) are edges (input and out edges included) number linked to $w^1$,$w^2$ respectively in WordNet.

We note that $sim_2(w \rightarrow w)$=0.5. And if $w^1 \neq w^2$ and N($w^2$)>>N($w^1$), $sim_2(w^2 \rightarrow w^1) \approx 0$, and $sim_2(w^1 \rightarrow w^2) \approx sim_1(w^1,w^2)$. So in asymmetrical similarities, a leopard is possible to be more similar with a cat than with itself.

3.2. **Collocation-based word similarities.** Unlike WordNet-based similarities, context-based similarities between words are based on the distributional hypothesis that "similar words appear in similar context". Context of words can be neighbor words, whole sentences or even whole paragraphs that contain them. In our work, we took collocations in two-word noun phrases as contexts of words.

[16] shows traditional context-based methods do not perform well in reflecting taxonomic semantic similarity. There are many efforts to bridge the gap between distributional similarities and semantic similarities [25-26]. Not like neighbor words in text, collocations reflect semantic but not distributional features. In noun-noun phrases, words' collocations always reflected some of their properties. For example as shown in Fig.2, when "paper" forms "paper cup", "paper flower", its properties "can be made some artifact" is reflected by "cup" and "flower". If collocations are similar, the words themselves are semantic similar in the reflected properties.

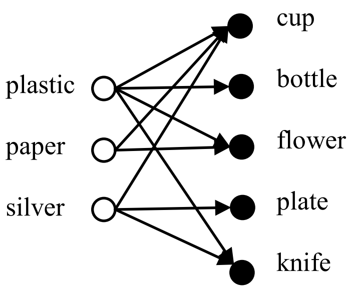

FIGURE 2. In noun-noun phrases, two words are semantic similar if their collocations are similar.

Not all words' properties can be reflected by collocations. Also taking "paper" for example, we could not find a noun-noun phrase to reflect its property of "opaque". That means when calculating two words' similarity using collocations, only parts of properties were involved. In this sense, we call collocation-based similarity as partial semantic similarity.

Partial semantic similarity is more valuable than traditional semantic similarity in our work. When "paper" and "glass" are calculated to be partial semantic similar in Fig.2, they were actually similar when they form noun-noun phrases using only part of their properties. Their other different but useless properties, such as "glass is transparent", "paper is opaque", will not stop them to be similar when judging whether "paper plate" and "glass plate" are similar or not.

### Symmetrical Collocation-based similarities

Collocation-based similarities calculating a typical SimRank problem[27]. Using SimRank idea, which is designed based on a simple and intuitive hypothesis that "two objects are similar if they are related to similar objects", collocation-based similarities between words in noun-noun phrases can be calculated by Eq.(3).

$$sim_3(w^1,w^2)=\frac{\sum_{i=1}^{|C(w^1)|}\sum_{i=2}^{|C(w^2)|}\left(sim_3\left(C_i(w^1),C_j(w^2)\right)\times \mathrm{wgt}\left(C_i(w^1)\right)\times \mathrm{wgt}\left(C_j(w^2)\right)\right)}{\sum_{i=1}^{|C(w^1)|}\mathrm{wgt}(C_i(w^1))\times\sum_{j=1}^{|C(w^2)|}\mathrm{wgt}\left(C_j(w^2)\right)} \quad (3)$$

In Eq.(3), we added weights of words on the standard SimRank formula. C($w$) is $w$'s collocations, $C_i(w)$ is the $i_{th}$ word in C($w$). $wgt(C_i(w))$ is tf-idf like weight of one collocation $C_i(w)$ in all collocations C($w$). If $w$ appeared more frequent in $C_i(w)$ or less frequent in all phrases, it has more weight as evaluated in Eq.(4).

$$\mathrm{wgt}(C_i(w))=\frac{|\{w\,C_i(w)\}|}{|\{*C_i(w)\}|}\times log\frac{|T_m|}{|\{*C_i(w)\}|} \quad (4)$$

Where $T_m$ is the multi-set, in which duplicate items can exist, obtained from corpus. After duplicate phrases are removed, it becomes T. $\{wC_i(w)\}$ is frequency of phrases $wC_i(w)$ and $wC_i(w)$ in $T_m$, and $*C_i(w)$ is all the phrases contains $C_i(w)$.

Like standard SimRank formula, Eq.(3) is recursive. In the beginning, two words have a similarity of 1 if they are absolute equal or 0 otherwise.

### Asymmetrical Collocation-based similarities

By collocations, we could quantify salience of words too. If a word is more salient, it is used more frequent and then created more phrases by people. So, we assumed that more collocations a word has, more salient it is. Accordingly, asymmetrical collocation-based similarities were calculated by Eq.(5)

$$sim_4(w^1 \to w^2) = \frac{sim_3(w^1,w^2)}{1+{|C(w^1)|}/{|C(w^2)|}} = \frac{sim_3(w^1,w^2)\times|C(w^2)|}{|C(w^1)|+|C(w^2)|} \qquad (5)$$

4. **Mining Relations by Analogizing.** Based on the hypothesis that "similar phrases contain similar semantic relations", a phrase is *r*-positive or not depends on whether its similar phrases are *r*-positive or not.

In our work, we take minimal value of similarities between corresponding words as similarity between phrases. The reason for giving up average and maximal ones can be demonstrated by an example, that is, "paper cup" and "paper cutter". Maximal similarity between their corresponding words is *sim*(paper,paper), which will incorrectly deduce that they contain same relation. And average similarity between these two phrases, 0.5*(*sim*(paper,paper)+*sim*(cup,cutter)), is also very likely to larger than average similarity between a pair of real similar phrases like "paper cup" and "plastic bottle".

Given a train set S in which each phrase is labeled to be *r*-positive or not, for a test phrases *t*, the probability of *t* to be *r*-positive is the possibility of its similar phrases in S are *r*-positive as a whole. We can compute the probability like Eq.(6).

$$\text{prob}(f(t) = true) = \frac{\sum_{t'\in P(t)} sim(t,t')}{\sum_{t'\in P(t)} sim(t,t') + \frac{|P(t)|}{|N(t)|}\times\sum_{t'\in N(t)} sim(t,t')} \qquad (6)$$

Where P(*t*) is *t*'s similar and *r*-positive phrases in train set S, and N(*t*) is similar but *r*-negative phrases in S. And *t*'s similar phrases are the ones with a similarity greater than a predefined threshold $\theta$ in S. Formally, $P(t)=\{t'|t'\in S\wedge sim(t,t')>\theta\wedge f(t')=true\}$, and $N(t)=\{t'|t'\in S\wedge sim(t,t')> \theta\wedge f(t')=false\}$.

We call S *seed phrases* or *seeds* because they can gather more *r*-positive phrases. S can be automatically obtained from raw corpus using lexico-syntactic patterns. Lexico-syntactic patterns like "$np_1$ is a [kind|sort|type] of $np_2$" or "$np_1$ such as $np_2$ and …" can extract relations from sentences[4]. Using a group of carefully designed patterns, a quantity of relations with very high precision can be obtained if low recall is tolerated, through the simple idea that only the relations that satisfied enough patterns are accepted[6]. Then the intersection of the result and phrases set T can be seen as S.

In Eq.(6), if no similar phrases are found in S, we cannot judge *t*. So if S is not big enough, there are lots of phrases cannot be analogized. In our method, the train set will grow with the analogizing process going like a snowball. Very reliable new found *r*-positive phrases are added to S and are utilized to analyze more phrases, and this procedure repeat until no more need seeds added.

The reason to merge new result to seeds is that similarities defined in Eq.(1-4) are not transitive. For instance, if *sim*("Europe", "continent")>$\theta$ and *sim*("continent", "ocean")>$\theta$, we cannot conclude *sim*("Eruope", "ocean")>$\theta$. Path might become longer and longer with the transitive operation in WordNet-based similarities, and in collocation-based similarities, common collocations are changed between any pair of words. So the new added seeds can be similar with some new phrases which the old ones are not.

In the end, phrases whose probabilities to be *r*-positive ones greater than a predefined threshold $\lambda$ are accepted to contain relation *r*.

## 5. **Experimental Result and Discussion.**

5.1. **Relation Specified and Evaluation Criteria.** In our experiment, we specified part–whole (or hol/meronymy) relation as *r*. Part-whole is considered a fundamental ontological relation since the atomists[5]. Based on psycholinguistic experiments and the way in which the parts contribute to the structure of the wholes, [28] determines six types of sub-relations. If phrase $t=w^1w^2$ contains a part-whole relation, it might be PART-OF($w^1$,$w^2$) or PART-OF($w^2$,$w^1$). We only focus on the former in the experiment, that is $f(t)=true$ iff PART-OF($w^1$,$w^2$) stands. For example, *f*("paper cup")=*true*.

We adopt precision and recall as the evaluation criteria as shown in Table 1.

TABLE 1. Precision and recall: precision=|A|/|B|, recall=|A|/|C|

| phrases in test set | set name |
|---|---|
| *r*-positive phrases mined automatically | A |
| all phrases mind automatically | B |
| *r*-positive phrases labeled manually | C |

5.2. **Performance.** We extracted 1,648,574 Chinese noun-noun noun phrases from raw corpus using syntax and semantic patterns described in [29] which a precision of 96.5%. Using lexico-syntactic methods described in [5], 3,955 in T are verified to be part-whole relations positive with an average precision of 83.3%, and be regarded as seeds S. Seeds partition in T is 0.24%. According to manual counting in [30], there is about 3.14% part-whole positive phrases in all Chinese noun-noun phrases. So seeds partition in all part-whole positive phrases of T is 7.40%, which demonstrate that majority of part-whole relations in phrases cannot be extracted by lexico-syntactic patterns.

We set $\lambda$=0.9 in experiment, which means we accepted *t* as a part-whole positive phrase if $prob(f(t)=true)>0.9$. Performances using four similarities calculating methods are shown in Fig.3.

We remind that $\theta$ is the minimal similarity that we regard two phrases as a similar pair in Eq.(6). From Fig.3, we can tell that precisions are all high and not influenced much by $\theta$. This is because we only keep the phrases whose probabilities to be part-whole positive degree greater than 0.9. Of course, when $\theta$ becomes higher and higher, seeds used for analogizing in Eq.(6) become more and more similar with the test phrase, and then precision can be further insured.

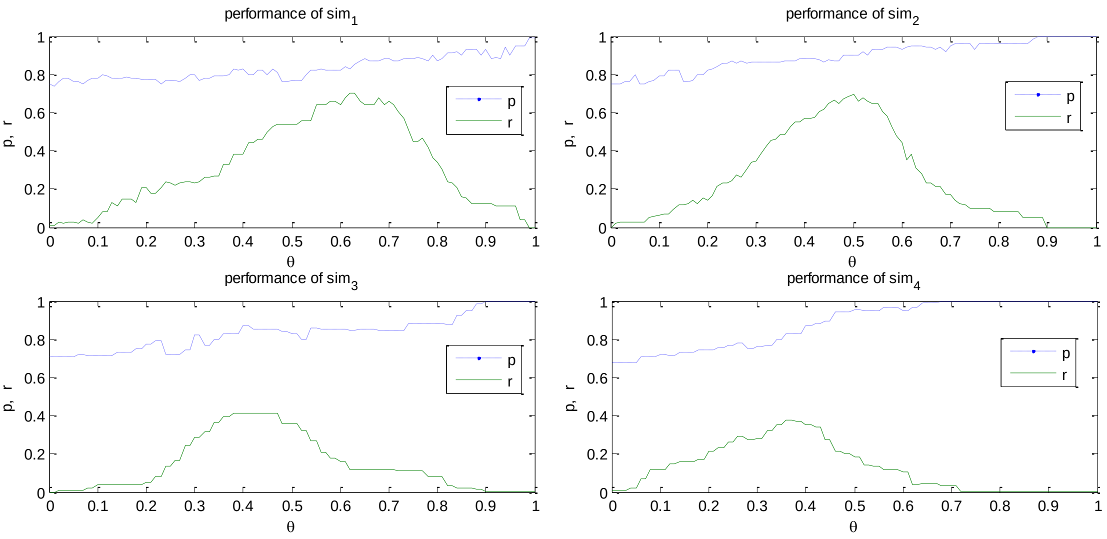

FIGURE 2. System performance under four words similarities calculating methods.

Recalls depend on $\theta$ heavily. Fig. 4 gives the average probability for all part-whole positive phrases in S. From it, we can see that when $\theta$ is very low, average probabilities is close to 0.5 because the similar phrases used in Eq.(6) are actually not similar at all. So there are few phrases which probabilities greater than 0.9 and be accepted. On the other side, when $\theta$ is too high, no similar seeds can be found and then the recall is also very low.

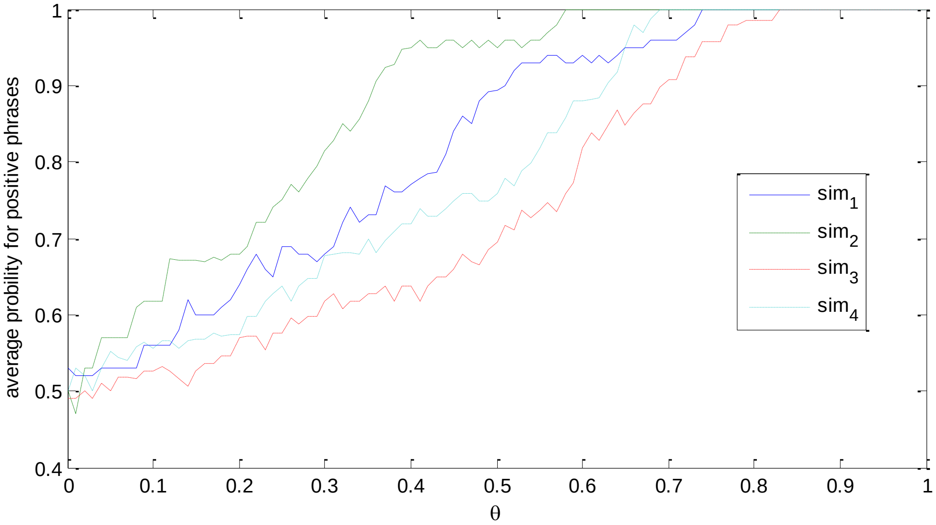

FIGURE 3. Average probability to be part-whole positive phrases in S.

The best result shown in Fig.3, which is occurred in asymmetrical similarities when $\theta$=0.51, achieves a recall of 0.69 and precision of 0.91. As a whole, WordNet-based similarities perform better than collocations-based similarities, and asymmetrical similarities perform better than symmetrical ones. Although demanding tremendous time and efforts, a well-organized ontology like WordNet helps a lot when calculating semantic similarities. Of course, the limitation of words coverage will restrict its usage.

Like traditional context-based similarities measuring approaches, collocations-based similarities often deduce relative low similarities because words' collocations contain lots of noises. Errors will be analyzed in detail in the next section.

5.3. **Errors Analysis and Discussion.** Through observing the results, we concluded that typical errors generated from two aspects. The major one is incorrect similarities calculated automatically. Taking "奶牛/cow 棚子/shed" for example, if "奶牛/cow" is wrongly calculated to be similar with "木材/wood" and "棚子/shed" is correctly measured to be similar with "房子/house", "奶牛/cow 棚子/shed" might be wrongly accepted to be part-whole positive phrase if "木材/wood 房子/house" is a seed.

In collocation-based similarities, "cow" is very likely to be measured to be similar with "paper". Below are the initial ten highest weighted collocations of the two words.

- C(奶牛/cow)=产业/industry，档案/document，工业/industry，公司/company，等级/level，基地/base，胚胎/fetus，品质/quality，品种/ brand，市场/market
- C(木材/wood)=产品/production，工业/industry，表面/surface，产业/industry，公司/company，产地/region,等级/level，品种/brand，市场/market，规格/standard

There are 6 shared collocations for the two words. So their collocation-based similarities is very high (0.58 in our experiment). About 98% errors are caused by this in the result. We have already used tf-idf weights to handle this problem in Eq.(4), more efforts are need in order to get better result.

Errors in WordNet-based similarities are also exists because the automatically translated Chinese WordNet contains many errors as described in [20]. Besides, polysemes also cause errors. For example, word "门槛" in Chinese has two senses, one is "a door" and the other is "a qualification". "大学/college 门槛/qualification" (the qualification to enter a college) is possible similar with "house door" if we use the largest similarities in WordNet between their corresponding concepts, and then incorrectly considered to contain a part-whole relation. Word sense disambiguation is needed for a better result in the future.

The second kind of errors are because the analogizing hypothesis of the proposed method. Some phrases, which are similar with some other phrases, actually do not contain similar relations. There are about 2% instances in the result. An example is "玻璃/glass 刀/knife", which is similar with "paper cup", "plastic plate", and "silver knife", but its meaning is actually "a knife which can cut glasses".

In this work, we did not consider the context of phrases. Some noun-noun phrases themselves are ambiguities. For example, "家具/furniture 装饰/decoration" can be explained to "the furniture is a part of the decoration", or "the decoration of the furniture". Properties of semantic relations are not touched neither. We do not distinguish between situations when whole objects consist of parts that are always present, or parts that are only sometimes present.

6. **Conclusion and Future Work.** In this paper, we proposed an automatic method to mine semantic relations from noun-noun phrases. The method is based on the hypothesis that similar phrases contain similar relations. Measuring similarities between phrases is the basic of the method. Four kinds of novel similarities were put forward and compared in this paper. Using a small group of seeds automatically obtained by traditional sentences-oriented approaches, an analogizing based algorithm which considered both positive and negative instances was presented to mine relations from phrases.

The proposed approach did not need manual efforts, and achieved an acceptable result in the part-whole relations. Although only focusing in one specified relation in experiment, the method can be applied to other relations directly. Besides, it can also be applied to other kinds of phrases.

Future work focuses on applying the method to other relations and phrases.

7. **Acknowledgment.** This work is supported by National Natural Science Foundation of China under grant No. 60773059, 61035004, 61173063, National High Technology Research and Development Program (863 Program) of China under No.2007AA01Z325, and National Social Science Foundation of China under grant No. 10AYY003.

## REFERENCES

[1] Barbara R., Marti A. H., Charles J. F.: The Descent of Hierarchy, and Selection in Relational Semantics. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 247-254, University of Pennsylvania, 2002.

[2] Lapata, M.: The Disambiguation of Nominalizations. Computational Linguistics, 28, 357—388, 2002.

[3] Morris, J., Graeme H.: Non-classical Lexical Semantic Relations. In: Proceedings of the 4th Human Language Technology Conference of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics, Workshop on Computational Lexical Semantics, pp. 46–51, Boston, MA, 2004.

[4] Hearst M. A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: 14th International Conference on Computational Linguistics, pp. 539-545, 1992.

[5] Roxana G., Adriana B., Dan M.: Automatic Discovery of Part-Whole Relations. Computational Linguistics, 32, pp. 83-135, 2006.

[6] Lei, L. Research on Theories and Methods of Extracting Concept and Hyponymy Relations. A dissertation Submitted to Graduate School of the Chinese Academy of Sciences for the degree of Doctor of Philosophy. Beijing, China, 2007.

[7] Sadao K., Masaki M., Yasunori Y., Mitsunobu S., Makoto N.: Construction of Japanese nominal semantic dictionary using "a の b" phrases in corpora. In: COLING-ACL’98 workshop on the

Computational Treatment of Nominals, 1998.

[8] Barbara R., Marti H.: Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: EMNLP’01, pp. 84—89, 2001.

[9] Stephen T., Eduard H.: A taxonomy, dataset, and classifier for automatic noun compound interpretation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010.

[10] Sadao K., Yasuyuki S.: Semantic analysis of Japanese noun phrases: A new approach to dictionary-based understanding. In: 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 481—488, 1999.

[11] George A. M.: WordNet: A lexical database for English. Communications of the ACM, 38, pp. 39–41, 1995.

[12] Fintan J. C., Tony V.: Using WordNet to automatically deduce relations between words in noun-noun compounds. In: COLING/ACL 2006 Main Conference Poster Sessions, pp. 160—167, 2006.

[13] Moldovan D., Badulescu A., Tatu M., Antohe D., Girju R.: Models for the semantic classification of noun phrases. In: HLT-NAACL 2004, Workshop on Computational Lexical Semantics, pp. 60–67, Boston, Massachusetts, USA, 2004.

[14] Jinglei Z., Hui L., Ruzhan L.: Semantic Labeling of Compound Nominalization in Chinese. In: Workshop on A Broader Perspective on Multiword Expressions, pp. 73–80, 2007.

[15] Walid S. S.: Language, Logic and Ontology: Uncovering the Structure of Commonsense Knowledge, International Journal of Human-Computer Studies, 2007.

[16] Akira U., Daisuke S.: Word Vectors and Two Kinds of Similarity, In: Proceeding of COLING/ACL on Main conference poster sessions, pp. 858-865, 2006.

[17] Amos T.: Features of similarity. Psychological Review, volume 84, pp. 327—352, 1977.

[18] Piek V.: Eurowordnet: a multilingual database with lexical semantic networks. Dordrecht: Kluwer Academic Publishers, 1998.

[19] Altangere C., Ho-Seop C., Cheol-Young O., Hwa-Mook Y.: On the Evaluation of Korean WordNet. In: Proceedings of the 10th international conference on Text, speech and dialogue, pp.123–130, 2007.

[20] Shi W., Cungen C.: WNCT: a Method for Automatic Translation of WordNet Concepts into Chinese. Journal of Chinese Information Processing. 67, 64—72, 2009.

[21] Blettner M., Rada R., Mili H., Bicknell E.: Development and application of a metric on semantic nets. IEEE Transactions on Systems Management and Cybernetics, 19, 17—30, 1989.

[22] ZhibiaoW., Martha S. P.: Verb semantics and lexical selection. In: Proceedings of the 32th Annual Meeting on Association for Computational Linguistics, New Mexico, USA, pp. 133–138, 1994.

[23] Yuhua L., Zuhair A., David M.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering. 2003, Vol. 15, 2003.

[24] Dongqiang Y., David M.W.: Measuring semantic similarity in the taxonomy of WordNet, In: ACM International Conference Proceeding Series, Proceedings of the Twenty-eighth Australasian conference on Computer Science, 315-322, 2005.

[25] James R. C.: From Distributional to Semantic Similarity, a dissertation submitted to University of Edinburgh for the degree of Doctor of Philosophy, 2003.

[26] Shi W., Cungen C., Ya-nan C., Han L., Xinyu C.: Measuring Taxonomic Similarity between Words Using Restrictive Context Matrices. In: 5th International Conference on Fuzzy Systems and

Knowledge Discovery, pp. 193-197, 2008.

[27] Glen J., Jennifer W.: SimRank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 538–543, New York, NY, USA, 2002.

[28] Winston, M., Chaffin, R., Herrmann, D.: A taxonomy of part-whole relations. Cognitive Science, 11:417-444, 1987.

[29] Shi W., Ya-nan C., Xinyu C., Cungen C.: Learning Concepts from Text Based on the Inner-Constructive Model. In: 2nd International Conference on Knowledge Science, Engineering and Management, pp.255-266, 2007.

[30] Shi W.: Research on Chinese Entity Names Recognition and Semantic Analysis. A dissertation Submitted to Graduate School of the Chinese Academy of Sciences for the degree of Doctor of Philosophy. Beijing, China, 2009.