Unsupervised Chinese Verb Metaphor Recognition Based on Selectional Preferences*

Yuxiang Jia, Shiwen Yu

Institute of Computational Linguistics, Peking University, Beijing, China, 100871
{yxjia, yusw}@pku.edu.cn

Abstract. Metaphors are pervasive in human language and developing methods to recognize and deal with metaphors is an indispensable task in Natural Language Processing (NLP). This paper proposes an unsupervised method to recognize metaphors from real texts. Firstly, source domain candidates are determined based on automatically acquired selectional preferences. And then metaphors are recognized with the source domain knowledge. Experiment results show that this unsupervised method outperforms the baseline by a great improvement. In addition, the source domain knowledge can also be used for metaphor comprehension.

Keywords: Selectional Preference, Source Domain, Metaphor Recognition, Concept Concreteness

1. Introduction

Conceptual Metaphor Theory (Lakoff and Johnson, 1980) considers metaphor as a mapping from the concrete source domain to the abstract target domain. Abstractions and enormously complex situations are routinely understood via metaphors. Metaphorical expressions are pervasive in human languages and must be treated for Natural Language Understanding (NLU) (Carbonell, 1982). As an important figure of speech, metaphor processing has interesting applications in many Natural Language Processing (NLP) tasks like machine translation, paraphrasing, information retrieval and question answering.

Metaphor processing can be divided into three tasks, recognition, comprehension and generation, among which recognition is the basic step. Metaphor recognition is to decide whether a sentence contains metaphorical expressions (a word, phrase or the whole sentence). This paper focuses on verb metaphor, to decide whether a verb is in metaphorical usage or literal usage. In selectional preference violation view, a satisfied preference indicates a literal semantic relation, while a violated preference indicates a metaphorical one. Take the following two sentences as examples.

(1) 农民　在　精心　培植　幼苗。
Nong2min2　*zai4*　*jing1xin1*　*pei2zhi2*　*you4miao3*
Farmer　at　carefully　cultivate　young plants
“Farmers are cultivating young plants carefully.”

* This research is funded by National Basic Research Program of China (No.2004CB318102). The authors are grateful to the two anonymous reviewers for their helpful comments and suggestions.

(2) 我们　要　大力　培植　人才。
Wo3men2　yao4　da4li4　pei2zhi2　ren2cai2
We　should　devote great effort　train　talents
“We should devote great effort to train talents.”

Sentence 1 is a literal usage while sentence 2 is a metaphorical one. The fact that literally 培植 ‘cultivate’ requires the object to denote some plants suggests that selectional preferences offer a cue to the presence of a metaphor. But the selectional preferences automatically induced by conventional computational models may not reflect semantics in the literal usage. On the other hand, concept concreteness or abstractness is an important indicator of literal usage, where concrete concepts usually indicate literal usage while abstract concepts correspond to non-literal usage. This paper makes use of concept concreteness based on automatically acquired selectional preferences for verb metaphor recognition.

Though metaphorical usage could be considered as a different sense of the target word, but when performing inference, it is beneficial to differentiate literal usage from metaphorical usage, because they share inferential structure. For example, the aspectual structure of 培植 ‘cultivate’ is the same in either domain whether it is literal or metaphorical. Further, this sharing of inferential structure between the source and target domains simplifies the representational mechanisms used for inference making it easier to build the world models necessary for knowledge-intensive tasks like question answering.

The rest of this paper is organized as follows. Section 2 is a review of related work. Section 3 describes the details of selectional preferences acquisition. Section 4 shows the method of source domain determination based on selectional preferences. Section 5 uses this source domain knowledge for metaphor recognition. Experiments and conclusions are given in section 6 and 7 respectively.

2. Related Work

Previous work on automatic metaphor recognition using selectional preferences idea includes (Martin, 1990), (Fass, 1991), (Mason, 2004) and (Krishnakumaran and Zhu, 2007). (Martin, 1990) detects metaphors by comparing new sentences with an empirically collected metaphor knowledge base and gives some interpretation of metaphorical sentences. (Fass, 1991) uses collative semantics to identify metaphors and distinguish metaphor from metonymy. But they both require hand-coded knowledge bases and thus have limited coverage.

(Mason, 2004) develops a corpus-based system CorMet for discovering metaphorical mappings between concepts. It finds selectional preferences of given verbs from automatically compiled domain-specific corpora, and then identifies metaphorical mappings between concepts in two domains based on differences in selectional preferences. (Krishnakumaran and Zhu, 2007) uses lexical resources like WordNet and bigram counts generated from a large scale corpus to classify sentences into metaphorical or normal usages. It does not compute selectional preferences explicitly and the bigram counts omit grammatical relations.

Later researches treat metaphor recognition task as a classification problem between normal and metaphorical usage. (Gedigian et al., 2006) uses a maximum entropy classifier to identify metaphors and takes verb arguments as features. (Wang et al., 2006) also uses a maximum entropy approach to recognize Chinese noun phrase metaphors. However, both need manually annotated corpus to train the classifier. In order to reduce manual work on annotation, (Birke and Sarkar, 2006) use a clustering approach with a smaller seed corpus to classify verb usages.

One advantage of selectional preferences based method is that it does not need training. One thing that sets our work apart is that all previous selectional preferences based methods do not make use of concept concreteness information. In contrast, we use it and show that it is effective information for metaphor processing.

3. Selectional Preference Acquisition

The automatic corpus-based induction of selectional preferences was first proposed by (Resnik, 1993). All later approaches have followed the same two-step procedure, first collecting argument head words from a corpus, then generalizing to other similar words. They are different mainly in the generalization step, some using manual semantic taxonomy like WordNet, while others using clustering methods.

Different from previous approaches, the first step in this approach is based on grammatical collocations. It makes use of various statistical measures for computing collocations or combination of some of them, not just word frequency used in previous approaches. For generalization, a semantic lexicon containing synonym and hypernym relations is employed.

3.1. Grammatical Collocation

Grammatical collocation means that the target word and its collocation are in a certain grammatical relation, such as subject-verb, verb-object or modifier-noun. In order to obtain grammatical collocations for the target word, this paper uses Sketch Engine (Kigarriff and Tugwell, 2001), a query system extracting collocations of different grammatical relations from a large scale corpus.

Collocations are sorted in descending order according to the salience value, which is estimated as the product of Mutual Information and log frequency. However, (Kilgarriff and Tugwell, 2001) modify the Mutual Information value by considering of the overall frequency of the grammatical relation as compared to other relations. The purpose of doing so is to avoid cases of low frequency collocations such as those which occur once but have high mutual information values because it is the only time they appear together with the target word. Therefore, the salience value is a reliable calculator instead of the frequency value.

The corpus for grammatical collocation extraction is the Simple Chinese Gigaword corpus, which has 706,427,624 tokens. The input parameters for Sketch Engine are as follows: the minimum frequency is 5; the minimum salience value is 0.0; the maximum number of items in a grammatical relation is 999, which is the upper bound due to licensing limitation.

As an example, table1 shows the top 20 collocations of the target verb 培植 pei2zhi2 ‘cultivate’ in the verb-object relation.

Table 1: Top 20 collocations, object of 培植 ‘cultivate’

Collocation	Frequency	Salience	Collocation	Frequency	Salience
人才	186	42.33	盆景	8	22.15
税源	29	39.92	幼苗	7	21.71
财源	55	38.18	新秀	12	20.57
干鱼	5	30.16	木耳	6	20.56
草坪	17	28.24	生长点	5	20.19
后进	13	27.58	人材	6	19.7
产业	86	24.29	蘑菇	6	19.4
新人	18	24.26	接班人	8	19.4
球员	36	23.91	细胞	16	18.17
增长点	12	23.69	势力	15	17.29

3.2. Semantic Mapping

Collocation words need to be generalized into semantic level to reflect the semantic preferences of the target word. A Chinese semantic lexicon named TongYiCiCiLin is used for this purpose. In the lexicon, about 80,000 words are arranged into 5-level tree structures (see figure1) according to semantic relations like synonym and hypernym. In the tree structure, the bottom level is called Atomic Word Group Level, where a node represents a synonym set. The parent node is the hypernym of the children. In total, 12 root nodes partition all words into 12 super classes, and the lower nodes further partition words into more detailed classes. The super classes include Human, Substance, Time and Space, Abstraction, Features, Motions, Psychological Activity, Activity, etc.

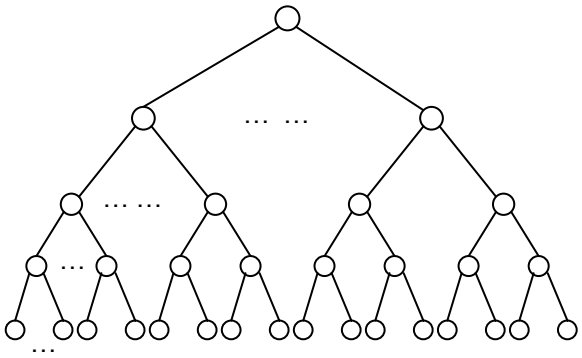

Figure1: The 5-level tree structure of TongYiCiCiLin

With the tree structure, collocations can be mapped to different semantic levels as required. After semantic mapping, collocations are grouped and semantic classes are sorted according to the number of collocations they contain. The sequence of semantic classes reflects the selectional preferences of the target word. The more collocations a semantic class contains, the more it is preferred. Table2 shows the top 10 semantic classes of level 2 of the target word 培植 'cultivate'. The first column is the semantic class ID in TongYiCiCiLin. The second column is the semantic class name. Column 3 and column 4 respectively show the number of collocations and collocations themselves of this semantic class.

Table 2: The top 10 semantic classes of 培植 'cultivate'

SCID*	SCName*	#of collocations	Collocations
Al	才识 'ability and insight'	6	人才 新秀 人材 骨干 艺术家 好手
Bh	植物 'plant'	6	幼苗 木耳 蘑菇 兰花 花卉 蔬菜
Db	事理 'reason and logic'	5	税源 财源 货源 资源 办法
Di	社会 政法 'society, politics and law'	5	工业 党 工作 地方 组织
Dd	性能 'performance'	5	实力 项目 方面 地方 组织
Ae	职业 'profession'	4	球员 选手 干部 厂商
Ba	统称 'general terms of substance'	4	农产品 产品 资源 植物
Cb	空间 'space'	3	生长点 方面 地方
Bk	全身 'body'	3	细胞 皮肤 骨干
Da	事情 境况 'event'	3	势力 过程 信息

*SCID=Semantic Class ID, SCName=Semantic Class Name.

4. Source Domain Determination

Semantic classes acquired in the last section need to be refined for metaphor processing. Some preferred semantic classes may denote metaphorical usage. For example, in table2, semantic class Al, ability and insight, is used metaphorically as the object of the target word 培植 'cultivate'. Only source domain candidates, semantic classes denoting literal usage, are useful knowledge for metaphor recognition and comprehension (Chung and Ahrens, 2006).

Usually, a concrete concept is used as the source domain while an abstract concept as the target domain. So the concept concreteness or abstractness is useful to determine source domains. Information of concept concreteness can be found in TongYiCiCiLin, where class Substance is concrete while class Abstraction is abstract.

We choose all concrete concepts in the top N (N=10 by default) semantic classes as the source domain candidates, and the choosing method is flexible. If no concrete concepts exist, the first semantic class is considered as the source domain candidate. The most preferred source domain candidate is considered as the real source domain when a metaphor occurs. The blocked lines in table3 show the source domain candidates of the target verb 培植 'cultivate' and the semantic class 'Bh' is the most preferred source domain, which agree well with human judgment. The semantic class ID beginning with 'B' denotes concrete class Substance.

Table 3: Source domain candidates of 培植 'cultivate'

SDC*	SCID*	#of collocations	Collocations
No	Al	6	人才 新秀 人材 骨干 艺术家 好手
Yes	**Bh**	**6**	**幼苗 木耳 蘑菇 兰花 花卉 蔬菜**
No	Db	5	税源 财源 货源 资源 办法
No	Di	5	工业 党 工作 地方 组织
No	Dd	5	实力 项目 方面 地方 组织
No	Ae	4	球员 选手 干部 厂商
Yes	**Ba**	**4**	**农产品 产品 资源 植物**
No	Cb	3	生长点 方面 地方
Yes	**Bk**	**3**	**细胞 皮肤 骨干**
No	Da	3	势力 过程 信息

*SDC=Source Domain Candidate, SCID=Semantic Class ID.

5. Recognition Algorithm

After the source domain candidates are determined, whether the target verb is literally or metaphorically used can be decided. If the object or the subject of the verb belongs to the source domain candidates, then it is literal usage; otherwise, it is metaphorical usage.

For example, the source domain candidates of 培植 'cultivate' are {Bh, Ba, Bk} as shown in table3. In 培植幼苗 'cultivate young plants', the object 幼苗 'young plants' belongs to semantic class Bh. So this is a literal expression. However, in 培植人才 'train talents', the object 人才 'talents' does not belong to Bh, Ba or Bk. So this is a metaphorical expression.

The pseudo code for the unsupervised metaphor recognition is as follows:

1. Parse the sentence and obtain object or subject headword of the verb.

2. Search the headword in source domain candidates. If found, then it is literal usage; else it is metaphorical usage.

6. Experiments

Experiments are set up to test performance of source domain determination and metaphor recognition.

6.1. Source Domain Determination

Twenty frequently metaphorically used verbs (see table4) are chosen to test source domain determination results. Measures are coverage and precision, which are defined in formula 1 and 2.

$$\text{Coverage}= \frac{\#\text{Verb whose real source domain occurs in top N semantic classes}}{\#\text{All verbs}} \quad (1)$$

$$\text{Precision}= \frac{\#\text{Verb whose real source domain is correctly determined}}{\#\text{All verbs}} \quad (2)$$

Table 5 shows the performance of source domain determination. As can be seen, only 7 verbs out of 20 have the most preferred semantic class as the real source domain, which indicates the necessity to introduce the conceptual concreteness information. 17 verbs have their real source domains occur in the top 5 preferred semantic classes, and 16 ones are correctly determined. All real source domains are covered in the top 10 semantic classes, and 17 ones are correctly found, with a precision of 85%. Errors occur when some concrete semantic classes are more preferred than the real source domains.

Table 4: 20 metaphorically used verbs

泛滥 fan4lan4	搁浅 ge1qian3	流失 liu2shi1	起飞 qi3fei1	起伏 qi3fu2
overflow	run aground	be washed away	take off	rise and fall
倾斜 qing1xie2	燃烧 ran2shao1	渗透 shen4tou4	瘫痪 tan1huan4	滑坡 hua2po1
slope	burn	permeate	paralyze	landslide
编织 bian1zhi1	点燃 dian3ran2	兜售 dou1shou4	兑现 dui4xian4	腐蚀 fu3shi2
weave	cause to burn	peddle	cash	corrode
解剖 jie1pou1	培植 pei2zhi2	提炼 ti2lian4	消化 xiao1hua4	净化 jing4hua4
dissect	cultivate	refine	digest	purify

Table 5: Source domain determination performance

Top N semantic classes	1	5	10
Coverage	7/20	17/20	20/20
Precision	7/20	16/20	17/20

6.2. Metaphor Recognition

Ten out of the twenty verbs in table4 are chosen to test the recognition performance. For each verb, about 40 sentences are extracted from People's Daily corpus and annotated as literal usage or metaphorical usage. The real usage distribution is shown in table6. As can be seen, 270 out of 413 samples are metaphorical ones, which account for 65.38%.

Table 6: Distribution of metaphorical usages for 10 verbs

Word	#Sample	#Metaphorical	#Literal
bian1zhi1	27	23	4

dian3ran2	33	10	23
jie3pou1	33	27	6
pei2zhi2	44	37	7
ti2lian4	38	26	12
fan4lan4	55	50	5
ge1qian3	45	31	14
qi3fei1	48	7	41
qi3fu2	40	23	17
tan1huan4	50	36	14
Total	413	270	143

Source domain candidates are checked and argument headwords of verbs are extracted manually to remove noises introduced by these steps, so that the capability of this recognition method can be examined given correct knowledge. Totally automatic experiments will be conducted in the near future. Measures of performance are defined as follows in formula 3 to 6.

$$\text{Precision}=\frac{\#\text{Correctly recognized metaphorical samples}}{\#\text{Recognized metaphorical samples}} \quad (3)$$

$$\text{Recall}=\frac{\#\text{Correctly recognized metaphorical samples}}{\#\text{All metaphorical samples}} \quad (4)$$

$$\text{F-measure}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (5)$$

$$\text{Accuracy}=\frac{\#\text{Correctly classified samples}}{\#\text{All samples}} \quad (6)$$

Table 7: Recognition performance

	Precision	Recall	F-measure	Accuracy
Baseline	65.38%	100%	79.07%	65.38%
Source domain	78.95%	100%	88.24%	82.57%
Source domain candidates	86.82%	100%	92.95%	90.07%

Three experiments are carried out (see table7). The baseline assumes that all samples are metaphorical usages and the F-measure is 79.07%. Experiment two only uses the most preferred source domain in the top 10 semantic classes as the source domain knowledge and achieves F-measure of 88.24%. Experiment three uses all source domain candidates in the top 10 semantic classes and the F-measure improves to 92.95%. The recognition method tries to removes recognized literal usages and leaves all others as metaphorical usages, so it always has high recall values.

7. Conclusions

This paper proposes an unsupervised metaphor recognition method based on selectional preferences. Different from other selectional preferences based methods, this approach utilizes concept concreteness information. Firstly, selectional preferences are extracted from a large scale corpus. Then source domain candidates are determined based on the acquired selectional preferences and concept concreteness information in a lexicon. Finally, source domain candidates are used for metaphor recognition and good performance is achieved. In addition, source domain knowledge is also helpful for metaphor comprehension.

More extensive experiments will be carried out to test the effectiveness of this approach. For comparison, supervised and semi-supervised classification methods will be examined.

Contextual information is useful for metaphor recognition, so more contextual information will be exploited to improve the method proposed in this paper.

References

Birke, J. and A. Sarkar. 2006. A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 329-336.

Carbonell, J.G. 1982. Metaphor: An Inescapable Phenomenon in Natural Language Comprehension. In W.Lehnert and M.Ringle eds., *Strategies for Natural Language Processing*, pp. 415-434, Hillsdale, N.J.: Lawrence Erlbaum.

Chung, Siaw-Fong and Kathleen Ahrens. 2006. Source Domain Determination: WordNet-SUMO and Collocation. *Proceedings of the 2nd International Conference of the German Cognitive Linguistics Association*, pp. 1-4.

Fass, D. 1991. met*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics*, 17(1), 49-90.

Gedigian, M., J. Bryant, S. Narayannan and B. Ciric. 2006. Catching Metaphors. *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pp. 41-48.

Kilgarriff, A. and D. Tugwell. 2001. Word Sketch: Extraction and Display of Significant Collocations for Lexicography. *Proceedings of the ACL Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation*, pp. 32-38.

Krishnakumaran, S. and X.J. Zhu. 2007. Hunting Elusive Metaphors Using Lexical Resources. *Proceedings of the Workshop on Computational approaches to Figurative Language*, pp. 13-20.

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live By*. Unversity of Chicago Press.

Martin, J. 1990. *Computational Model of Metaphor Interpretation*. San Diego: Academic Press.

Mason, Z.J. 2004. CorMet: A Computational, Corpus-based Conventional Metaphor Extraction System. *Computational Linguistics*, 30(1), 23-44.

Mei, J.J., Y.M. Zhu and Y.Q. Gao. 1983. *Tongyici Cilin*. Shanghai: Shanghai Cishu Press.

Resnik, P. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Wang, Z.M., H.F. Wang, H.M. Duan, S. Han and S.W. Yu. 2006. Chinese Noun Phrase Metaphor Recognition with Maximum Entropy Approach. *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 235-244.