2011 Seventh International Conference on Computational Intelligence and Security

Construction of a Bilingual Cognitive Property Knowledgebase

Bin Li[1,2] Haibo Kuang[2] Xiaohe Chen[2] Xuri Tang[3] Chen Chen[4]
1. State Key Lab for Novel Software Technology
Nanjing University
Nanjing, PR China
2. Research Center of Language and Informatics
Nanjing Normal University
Nanjing, PR China
3. College of Foreign Languages
Wuhan Textile University
Wuhan, PR China
4. Department of English Language and Literature
Graduate School of Shanghai International Studies University
Shanghai, PR China
libin.njnu@gmail.com, mchypocn@hotmail.com, {chenxiaohe5209, xrtang}@126.com

***Abstract*—Every language has its own culture background, thus it is difficult to translate or retrieve figurative expressions across languages. Based on the metaphoric cognition and feature analysis theory, we collect data from the web to construct the Chinese-English bilingual lexical cognitive property knowledgebase linked to "HowNet". By comparing the differences of the cognitive property, we get some answers to the core linguistic problem "is the metaphor universal?" Then, we put forward a novel method to gain the metaphoric property of a word by its translation in another language. The paper lays a solid foundation for semantic analysis and computing of lexical cognitive property.**

Keywords-cognitive property; linguistic knowledgebase; metaphor; lexical semantics

I. INTRODUCTION

Every language is coded with its culture, and many words have the conceptualized meanings in different languages. But most of the cultural meanings are not described in dictionaries. For example, the noun "pig" means fat when referred to a man both in English and Chinese, but in English "pig" also means dirty while not in Chinese. This kind of meaning is traditionally taken as cultural meaning, salient property or metaphor property of a word(Hao 2010). From the perspective of cognitive linguistics(Lakoff 1980; Bowdle 2005), we call it "cognitive property" as it conceptualizes speakers' everyday cognitive feelings in a language community.

To compare the cognitive property of words across languages is very useful in machine translation, cross language retrieval and language teaching. Therefore, we construct the Chinese-English bilingual lexical cognitive property knowledgebase and make statistical analysis on the basis of the concept of metaphor cognition and the theory of word property analysis. We choose "HowNet" (Dong 2006) as the semantic lexicon. The construction of the Chinese-English bilingual lexical cognitive property knowledgebase provides the quantitative statistic for finding the similarities and differences of the lexical metaphor cognitive mechanism and comparative analysis of the lexical metaphor property across languages. The knowledgebase will facilitate the research of the lexical metaphor cognitive mechanism and semantic computing of the lexical metaphor.

II. RELATED WORK

To collect the cognitive property is definitely difficult. However, some scholars have found efficient methods. Kintsch(2000) collected the noun-adjective word pairs like "pig-fat" with the Latent Semantic Analysis(LSA) on large corpora. Roncero(2006) considers the similes which contain the specific metaphor property. Veale(2007) and Hao (2010) argue that there is an evolutionary path from simile mechanism to metaphor mechanism. They collect a large scale of English similes to construct the English lexical metaphor property knowledgebase, which contains items as "noun vehicle-adjective property", using search engine and WordNet[1]. In a similar way, Jia(2009) collects Chinese similes to construct the Chinese lexical metaphor property knowledgebase, which contains items as "noun vehicle-adjective property" as mentioned above, using Cilin[2] as the semantic resource.

All the previous works are done within mono-language. Therefore, we want to extend the research to multi-languages.

III. CONSTRUCTION OF THE LEXICAL COGNITIVE PROPERTY KNOWLEDGEBASE

We use the specific simile sentence to collect Chinese similes as Veale(2007) and Jia(2009) by querying the search engine and then construct the Chinese-English bilingual

1 http://wordnet.princeton.edu/
2 http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=112

978-0-7695-4584-4/11 $26.00 © 2011 IEEE
DOI 10.1109/CIS.2011.244

lexical cognitive property knowledgebase. The Chinese knowledgebase will also be compared with Veale's base.

A. Lexcial Sematic Resource: HowNet

HowNet [3] is a structured bilingual(Chinese/English) semantic resource. Different from WordNet, it defines a word's meaning by a set of structured semantic features, called "sememe". In HowNet(version 2007), there are about 2200 sememes, which are used to define 91000 Chinese words and 85000 English words. For example, in HowNet the noun 猪(pig) and 笨(stupid) are defined as follows.

猪-pig, noun：{livestock|牲畜}

笨-stupid, adjective：{foolish|愚}

The definition explains the word's semantic class, related events and domains, all using sememes. Comparing to WordNet, the sememes in HowNet are not isolated from each other but structured with different kinds of relations. All sememes are located in a hierarchy tree. It has its own definition and can inherit its father node's definition. For example, the sememe "livestock|牲畜" is not just a semantic label, but a semantic feature with full meaning and kinds of relation to other sememes(see below). So in HowNet, a word's meaning is represented by many sememes.

{entity|实体}
├{thing|万物} {entity|实体:{ExistAppear|存现:existent={~}}}
｜ ├{physical|物质} {thing|万物:HostOf={Appearance|外观},{perception|感知:content={~}}}
｜ ｜ ├{animate|生物} {physical|物质:HostOf={Age|年龄}, {alive|活着:experiencer={~}},{die|死:experiencer={~}},{metabolize|代谢:experiencer={~}},{reproduce|生殖:PatientProduct={~},agent={~}}}
｜ ｜ ｜ ├{AnimalHuman|动物} {animate|生物:HostOf={Sex|性别}, {AlterLocation|变空间位置:agent={~}},{StateMental|精神状态:experiencer={~}}}
｜ ｜ ｜ ｜ ├{human|人} {AnimalHuman|动物:HostOf={Ability|能力}{Name|姓名}{Wisdom|智慧},{speak|说:agent={~}},{think|思考:agent={~}}}
｜ ｜ ｜ ｜ └{animal|兽} {AnimalHuman|动物:{GetKnowledge|认知:adjunct={neg|否},agent={~}}}
｜ ｜ ｜ ｜ ├{beast|走兽} {animal|兽:modifier={wild|野生}}
｜ ｜ ｜ ｜ ├**{livestock|牲畜}** {animal|兽:MaterialOf={edible|食物}, modifier={domesticated|家养},{eat|吃:patient={~}},{foster|饲养:patient={~}}}

B. English Resource: Sardonicus

The English lexical metaphor property knowledgebase "sardonicus" [4] built by Veale(2007) has 74704 simile sentences and maps 3769 different adjective property to 9286 noun vehicles. The mapping relationships are classified to two categories: "factual" (like horse-strong) and "ironic" (like ant-strong).

Based on sardonicus, we manually filter out the simple comparison items and the error items, then 10411 "noun-adjective(n-a)" pairs and 3585 types of noun left. A noun vehicle has an average of 2.90 adjectives and 51% of all the vehicles have only 1 adjective, 15.7% have 2 adjectives, 13.3% have 3 adjectives, 20% have more than 4 adjectives.

Sardonicus is built on English data. In order to do the comparison across languages, we have to give the Chinese and English expressions of noun-adj pairs. When connected to HowNet, only 6083 n-a items are left in Sardonicus, because many words and phrases are not included in HowNet.

C. The Chinese Cognitive Property Knowledgebase

To find the similarities and differences between Chinese and English, we also conduct a search engine based method to collect "noun-adjective" word pairs. Using specific simile sentence "X像Y一样P"(which means "X is as P as Y") as Jia(2009), we query the Chinese search engine Baidu(www.baidu.com) for each noun and adjective in HowNet. Then 18205 tokens of vehicle-adj simile pairs are collected. Then we trim the error itmes and comparison items. As Veale(2007), we also manually classify the simile word pairs into two categories: Factual and Ironic. The difference is that, we reserve the frequencies of word pairs to take a further observation. As a result, the Chinese lexical cognitive property knowledgebase has 4002 n-a types and 1908 vehicles.

It can be observed that the distribution of the number of the lexical metaphor properties in Chinese database. 81% of all the vehicles have 1 adjective, 10.18% have 2 adjectives, 13.82% have 3 adjectives, 5% have more than 4 adjectives. Compared to the English base sardonicus, it is not difficult to find that the certainty of Chinese lexical metaphor property is stronger than English because of the different data scale and data sparse degree.

As mentioned above, we collect the frequencies of items in the Chinese base. The frequency is useful, which will help us get the most frequent simile expressions in Chinese(see table I).

TABLE I. TOP 5 MOST FREQUNET NOUN-ADJ PAIRS IN CHINESE BASE

ID	Noun-Adj Pair	Frequence
1	百度-方便（baidu-convenient）	390
2	花儿-美丽（flower-beautiful）	91
3	自由-美丽（freedom-beautiful）	89
4	风-自由（wind-free）	73
5	钢铁-硬（iron-hard）	65

IV. INTEGRATION OF THE CHINESE-ENGLISH BILINGUAL LEXICAL METAPHORS PROPERTY KNOWLEDGEBASE

After collecting the Chinese "noun-adjective" pairs, we got a new resource to compare with sardonicus. Table II shows the top 5 nouns with the largest number of adjectives in sardonicus. In contrast, table III gives the top 5 nouns in Chinese bank. The comparison is very interesting, as we come to know that the most frequent used vehicle nouns in English and Chinese are different.

TABLE II. TOP 5 MOST FREQUNET NOUNS IN SARDONICUS

ID	Noun	Adjectives	# of ADJs
1	rock	unconscious, stupid, dumb, firm	44
2	snake	naked, smooth, heartless, artful	26

3 http://www.keenage.com

4 http://afflatus.ucd.ie/sardonicus/tree.jsp

3	diamond	unique, precious, glorious, pure	25
4	cat	sensitive, lazy, mysterious, curious	24
5	mountain	steady, calm, strong, solid	24

TABLE III. TOP 5 MOST FREQUNET NOUNS IN CHINESE BASE

ID	Noun	Adjectives	# of ADJs
1	水(water)	清 clear,纯净 pure,软 soft	37
2	花儿 (flower)	漂亮 beautiful,纯洁 pure,甜蜜蜜 sweet	27
3	猪(pig)	可爱 lovely,笨 stupid,懒 lazy	24
4	天空(sky)	寂寞 lonely,纯洁 pure,高 high	24
5	男人(man)	坏 bad,自私 selfish,彪悍 strong	23

To answer the key question whether cognitive properties of a concept are universal in different languages, we consider making a preliminary integration of the base and constructing a new Chinese-English bilingual lexical cognitive property knowledgebase. As introduced above, the English and Chinese cognitive property bases have been linked to HowNet, in which every word's meaning is defined by a number of basic concepts(sememes). Thus the sememes can be used to see if the nouns of the same concept in different languages have the same cognitive properties.

By linking the items in Chinese and English, we have 1065 bilingual noun-adj pairs and 76 vehicles. It is not difficult to find that the cognitive properties of some vehicles are bilingually the same. Table IV shows the top 10 noun vehicles which having the largest number of cognitive properties. The first adjective's sememes are also given in order to incorporate the synonyms. The nouns "水晶 crystal", "花 flower" and "妈妈 mother" have the same cognitive properties conveyed by adjectives in both languages. Now, the answer to the key question is clear that some concepts in different languages do have the same or similar cognitive properties.

TABLE IV. TOP 10 MOST SIMILAR VEHICLES IN THE CHINESE-ENGLISH BILINGUAL BASE.

ID	Chs Noun	Eng Noun	ADJ	ADJ:Sememe
1	水晶	crystal	清,清澈,纯,纯净-pure；脆-clear	pure:{spotless|洁}
2	花	flower	新-fresh；甜-sweet；纯真-pure	sweet:{sweet|甜}
3	妈妈	mother	好-good；温柔,柔和-gentle	gentle:{gentle|柔}
4	蚂蚁	ant	慢-slow；渺小,小-tiny	small:{small|小}
5	蛋糕	cake	甜美,甜蜜-sweet,luscious	sweet:{sweet|甜}
6	糕点	cake	甜美,甜蜜-sweet,luscious	sweet:{sweet|甜}
7	糖	sugar	甜蜜,好吃-sweet,nice	sweet:{sweet|甜}
8	婴儿	baby	裸-bare,naked	naked:{ naked|赤裸}
9	海洋	ocean	宽广,大-broad	broad:{broad|广}
10	针	needle	锋利-sharp,incisive	sharp:{sharp|利}

On the other hand, we focus on the items which do not have the same cognitive properties. It is the differences of the cognitive results between the 2 languages. Table V and VI give the top 5 vehicles which have the largest numbers of the cognitive properties in one language knowledgebase but does not appear in the other language. We suggest that these cognitive properties are language dependent. However, this assumption may be too strong. Some of the properties may be not language dependent but be missed by search engine queries. In next section, we try to make some effort in collecting more cognitive properties by bilingual information.

TABLE V. TOP 5 DEPENDENT VIHICLE NOUNS IN ENGLISH.

ID	Vehicles	Properties（Factual/Ironic）
1	kitten	happy,ineffective,unworldly,…
2	statue	hard,deadly,stiff,…
3	puppy	lovable,sweet,gentle,cozy,…
4	snowflake	Intricate,natural,pure,…
5	tornado	capricious,deadly,violent,cute,…

TABLE VI. TOP 5 DEPENDENT VIHICLE NOUNS IN CHINESE.

ID	Vehicles	Properties（Factual/Ironic）
1	天气 (weather)	好，火热(hot),糟阴(bad),阴冷(cold),…
2	阳光(sun light)	健康(healthy),绝望(despair),透明(diaphaneity),…
3	心情(state of mind)	晴朗(sunny),乱(disordered),快乐(happy),…
4	牛奶(milk)	嫩(soft),白皙(white),美丽(beautiful),…
5	奥运 (Olympics)	快(fast),随意(easy),畅通(unblocked),…

V. MUTUAL-GAIN OF THE COGNITIVE PROPERTY ACROSS LANGUAGES

The study of mutual-gain of the metaphor property across languages is another task in construction of the cross-language cognitive property knowledgebase. As shown in table V and VI, many of the frequent nouns are not included in sardonicus or the Chinese base. Thus we put forward a metric for gaining more cognitive properties of a noun. The algorithm is quite simple: if a word pair "noun-adjective" in one language base is not seen in another language base, then the simile template(like "N is as ADJ as N") is used to extract the translation word pairs.

We select 10 vehicles in the English knowledgebase to gain Chinese "noun-adjective" pairs. The result is shown in table VII. The blacked items are the new-gained cognitive

properties. At first, the properties are less. And after gaining, the Chinese noun vehicles get more cognitive properties.

TABLE VII. MUTUAL-GAIN OF CHINESE NOUNS

ID	Noun Vehicles	English Property	Chinese Property after Gaining (word_frequency)
1	abacus 算盘	primitive	死板_1，坚硬_2
2	abattoir 屠宰场	bloody	性感_2，无奈_1，恶心_1
3	chef 厨师，大师傅，主厨	fastidious skilled **expert**	创新_1，**专业_2**，出色_2
4	tuna 金枪鱼	intriguing (奇妙)	被捕杀_10
5	torrent 湍流	swift concentrated	汹涌_3，奔放_1,**急_2**
6	waterfall 瀑布	natural dynamic magical **lovely** spectacular	飞泻_1，狂飙_2，义无反顾_1，**漂亮_3**
7	lynx 猞猁	fearless	浓密_1，神秘_1
8	loon 懒人	mad daft nutty crazy stupid	去设计_4，简单_2
9	map 地图	accurate precise orderly	纹_1，搜索_2，有层级_1
10	seal 海豹	wet smooth	蠕动_2，没有四肢_5,可爱_1

VI. CONCLUSION AND FUTURE WORK

In this paper, we construct the Chinese-English bilingual lexical cognitive property knowledgebase linked with the semantic resource HowNet. The lexical cognitive property and their comparison results describe the features of the lexical metaphors in multi-view of points across languages: some nouns have the universal cognitive properties while most nouns have the language dependent properties. The phenomenon can partly explain why machine translation is difficult, and the knowledgebase will be a good foundation for semantic processing in machine translation.

Our next step will continue to extend the knowledgebase in data scale and language types. Second, we want to make more detailed analysis on the data and the categorization procedure in different languages. Third, we will do some experiments in machine translation using the cognitive property knowledgebase to see if the base is useful in real applications of natural language processing.

ACKNOWLEDGMENT

This paper is supported in part by National Social Science Fund of China under contract 10CYY021, 11CYY030, State Key Lab. for Novel Software Technology under contract KFKT2011B03, Jiangsu PostDoc Fund under contract 1101065C, National Natural Science Fund of China under contract 60773173.

REFERENCES

[1] Bowdle, B. F. and Gentner, D. “The Career of Metaphor”. Psychological Review, vol. 112, 2005.

[2] Dong, Z. D. & Dong, Q. “HowNet and the Computaion of Meaning”. Singapore, World Scientific Press, 2006.

[3] Gcntner, D. & B. F. Bowdle. “Convention,Form,and Figurative Language Processing”. Metaphor and Sym-bol, vol. 16, 2001.

[4] Hao, Y. F. & Veale, T. “An Ironic Fist in a Vel-vet Glove:Creative Mis-Representation in the Construction of Ironic Similes”. Minds and Machines, Vol. 20, No. 4, 2010.

[5] Jia Y. X. and Yu S. W. “Instance-based Metaphor Comprehension and Generation”. Computer Science, vol. 36, no.3, pp.138-41, 2009.

[6] Kintsch, W. “Metaphor comprehension: A computational theory”. Psychonomic Bulletin & Review, vol.7, pp.257-66, 2000.

[7] Lakoff,G. & Johnson.M. “Metaphors We Live by” .Chicago:The University of Chicago Press, 1980.

[8] Roncero, C., Kennedy, J. M., and Smyth, R. “Similes on the internet have explanations”. Psychonomic Bulletin and Review, vol.13, no.1, 2006.

[9] Veale, T. & Hao, Y. F. “Learning to Understand Figurative Language: From Similes to Metaphors to Irony”. Proceedings of CogSci 2007, Nashville, USA, 2007.