Web Based Collection and Comparison of Cognitive Properties in English and Chinese

Bin Li[1,2] Jiajun Chen[1] Yingjie Zhang[1]
[1.] State Key Lab for Novel Software Technology
Nanjing University
[2.] Research Center of Language and Informatics
Nanjing Normal University
Nanjing, PR China
{lib;chenjj;zhangyj}@nlp.nju.edu.cn

Abstract

Cognitive properties of words are very useful in figurative language understanding, language acquisition and translation. To overcome the subjectivity and low efficiency in manual construction of such database, we propose a web-based method for automatic collection and analysis of cognitive properties. The method employs simile templates to query the search engines. With the help of a bilingual dictionary, the method is able to collect tens of thousands of "vehicle-adjective" items of high quality. Frequencies are then used to obtain the common and independent cognitive properties automatically. The method can be extended conveniently to other languages to construct multi-lingual cognitive property knowledgebase.

1 Introduction

Cognitive Linguistics focuses on the cognitive and metaphorical usage in language. For example, In English the "pig" is fat, dirty and lazy, etc. But it is not the case in other languages. As in Chinese, 猪 (pinyin: zhu, means pig) is fat, lazy and happy, but not dirty. Different cultural backgrounds lead to differences in everyday cognitive knowledge (Lakoff 1980). Therefore it is beneficial for literature translation, cross language retrieval and language acquisition to compare the cognitive properties of words across languages. Traditionally, this kind of knowledge is generally possessed by experienced translators. In this article, we propose a method to collect the knowledge from the web automatically. It also makes a comparison between the obtained results with a traditional bilingual dictionary.

2 Related Work

To collect the cognitive properties by hand is considered as both labour intensive and subjective. Therefore the researchers have sorted to corpus and search engine for help. Kintsch(2000) collects the noun-adjective pairs like "pig-fat" using the Latent Semantic Analysis(LSA) method on a large corpora. Roncero(2006) considers the simile sentences which contain the specific metaphor property like "as adjective as noun". Veale(2007) collects a large scale of English similes by querying the nouns and adjectives in WordNet from Google to construct the English lexical metaphor knowledgebase "sardonicus", which contains about 10,000 items of "noun vehicle-adjective property". In a similar way, Jia(2009) collects Chinese similes from Chinese search engine Baidu. A total number of about 20,000 "noun vehicle-adjective property" items were acquired.

Querying search engines is an efficient way to collect "noun-adjective" items. However, all the previous works are monolingual and do not use the frequencies of the items. Therefore, we want to extend the research to multi-languages and use frequency for the comparison of cognitive properties.

3 Construction of the Bilingual Cognitive Property Knowledgebase

Just like Veale(2007) and Jia(2009), we use specific simile templates to collect English and Chinese "noun vehicle-adjective property" items by querying the search engines and then construct the Chi-

nese-English bilingual lexical cognitive property knowledgebase.

The words in WordNet and HowNet are used for querying the search engines. For English, the adjectives in WordNet are used. For Chinese, the words are taken from HowNet.

3.1 Lexical Resources

WordNet 3.0 is a widely used lexical resource, which contains 21,479 adjectives and 117,798 nouns(Miller 1990). It supplies plenty words for collecting English similes.

HowNet is a structured Chinese-English lexical semantic resource(Dong 2006). Different from WordNet, it defines the meaning of a word by a set of structured semantic features, named "sememes". About 2200 sememes are used to define 95000 Chinese words and 85000 English words In HowNet(ver. 2007). For example, the noun 猪(pig) and 笨(stupid) are defined as follows.

猪-pig, noun：{livestock|牲畜}

笨-stupid, adjective：{foolish|愚}

3.2 English Item Collection

We used the 21,479 adjectives in WordNet to fill in the simile template "as ADJ as". When querying Google, 3 limitations are set in advanced search to refine the search results: exact phrase, English language and up to 100 results for each query. We do not use the nouns in WordNet, but the template will supply thousands of nouns where querying Google. Thus, a number of 585,300 types (1,054,982 tokens) of "as…as…" items are gathered from Google. To trim the great number of nonsense, noisy and erroneous items, Veale(2007) manually checks the returned results. It is accurate but takes too much time. We introduce a simple trick for the purpose, which uses the dictionary for filtering. Nouns and adjectives in HowNet are taken to filter the "noun-adjective" items. Then, 27,331 types (87,529 tokens) of "noun-adjective" items are left, covering 6,319 nouns and 4,100 adjectives. Table 1 gives the top 10 most frequent items with their frequencies.

The item "blood-red" is the most frequent one in English. The frequency can tell the salience of the cognitive properties of nouns. Nevertheless, the frequencies we got are not exactly the frequency of the items on the web. They only show the statistical situation in the collected items.

TABLE 1. Top10 most frequent vehicle-adjective items in English

ID	VEHICLE	ADJ	FREQ
1	blood	red	628
2	twilight	gay	466
3	grass	perennial	413
4	ice	cold	392
5	mustard	keen	385
6	snow	white	340
7	sea	boundless	314
8	feather	light	289
9	night	black	280
10	hell	mad	254

The frequency of "blood-red" is over 100, because it also occurs in returned results of other words. Ideally, it is better to use the simile template "as ADJ as NOUN" for the parings of 21,479 adjectives multiple 117,798 nouns. But the limitation of the frequency to query search engines makes it impossible to finish the collecting work within a short time.

3.3 Chinese Item Collection

For Chinese, there are more simile templates. Three templates "像(as)+NOUN+一样(same)", "像(as)+VERB+一样(same)", "像(as)+一样(same)+ADJ" are adopted and are filled with the 51020 nouns, 27901 verbs and 12252 adjectives from HowNet to query Baidu(www.baidu.com). Verbs are also considered, because some of them may function grammatically as nouns in English. For example, "呼吸(breath)" is a verb in Chinese, but it may serve as a noun phrase in certain contexts, and one of its cognitive properties extracted from Baidu is "自然(natural)". It tells people's experience in breathing. We submit 91173 queries to Baidu, with configurations set to 100 returned results for each query. Totally, 1,258,430 types (5,637,500 tokens) of "vehicle-adjective" items are gathered. Then, nouns and adjectives in HowNet are used to filter these items, leaving only 24,240 items. The web database of the Chinese filtered items is already available for search at `http://nlp.nju.edu.cn/lib/cog/ccb_nju.php`. Table 2 shows the top 10 most frequent items with their frequencies.

TABLE 2. Top10 most frequent vehicle-adjective items in Chinese

ID	VEHICLE	ADJ	FREQ
1	苹果 apple	时尚 fashionable	1445
2	呼吸 breath	自然 natural	758
3	晨曦 sun rise	朝气蓬勃 spirited	750
4	纸 paper	薄 thin	660
5	雨点 rain drop	密集 dense	557
6	自由 freedom	美丽 beautiful	543
7	雪 snow	白 white	521
8	花儿 flower	美丽 beautiful	497
9	妖精 spirit	温柔 gentle	466
10	大海 sea	深 deep	402

It is surprising to see that "apple" has taken the first place on the web media in China. And "snow-white" occurs in the top10 place in both languages. In next section, we will compare the cognitive properties based on the collection works done on Google and Baidu.

4 Bilingual Comparison

Previous sections have already done some comparison by showing the most frequent items in English and Chinese. In this section, we continue to find the common parts and differences in cognitive properties.

4.1 Common vehicles and properties

We can compare the common vehicles and properties in English and Chinese. By consulting HowNet, 3,106 types of bilingual "vehicle-property" items are gathered, including 1,500 English items and 2,254 Chinese items. They cover only about 10% of all items in each language.

Table 3 shows the top 10 most frequent bilingual items. We can see that people in different cultures share many same properties of things, such as "snow-white", "blood-red". However, the "fox-sly" is somewhat strange and interesting, for the animal is not as smart as man or monkey, but is considered sly. About 90% of the "vehicle-adjective" items do not have their corresponding items in the other language. But it does not necessarily mean that the two languages share few common parts. Too many words miss their translations only due to the size of the bilingual dictionary HowNet. For example, "snazzy" and "popular" are not translated to "时尚" or "时髦" in HowNet. Thus, "apple" does not appear in the bilingual common items. So a larger bilingual dictionary is necessary in further researches. However, no matter how large the dictionary is, it may still encounter the difficulty to find all the translation word pairs.

TABLE 3. Top10 most frequent vehicle-adjective pairs in English and Chinese

ENG VEHICLE	ENG ADJ	ENG FREQ	CHS VEHI-CLE	CHS ADJ	CHS FREQ
snow	white	340	雪	白	521
blood	red	628	血	红	227
paper	thin	132	纸	薄	660
ice	cold	392	冰	冷	256
feather	light	289	羽毛	轻	111
honey	sweet	55	蜜	甜	324
sea	bound-less	314	大海	广阔	63
steel	strong	64	钢铁	硬	194
fox	cunning	88	狐狸	狡猾	166
fox	sly	85	狐狸	狡猾	166

4.2 Dependent vehicles and properties

As can be seen below, the "vehicle-property" items depend on culture backgrounds.

TABLE 4. Top10 most frequent dependent vehicle -adjective pairs in English and Chinese

ENG VEH	ENG ADJ	ENG FRQ	CHS VEH	CHS ADJ	CHS FRQ
twilight	gay	466	苹果 apple	时尚 fashionable	1445
grass	peren-nial	413	呼吸 breath	自然 natural	758
mustard	keen	385	晨曦 sun rise	朝气蓬勃 spirited	750
hell	mad	323	雨点 rain drop	密集 dense	557
life	large	288	自由 freedom	美丽 beautiful	543
punch	pleased	254	妖精 spirit	温柔 gentle	466
beetroot	red	240	阳光 sunlight	灿烂 resplendent	386
hatter	mad	226	天神 deity	美丽 beautiful	341
school children	cruel	209	天使 angle	美丽 beautiful	337
moun-tain	im-mova-ble	100	裁判员 referee	狠 ruthless	300

Most of the items are dependent on their language and culture. Table 4 shows the top10 most frequent independent items in English and Chinese, But when a bilingual dictionary is used, some

items are wrong like“苹果-时尚”and “天使-美丽”, as HowNet does not give good translations. With the bilingual cognitive properties, we can see the cognitive property differences among languages in a quick and convenient fashion. It will supply useful information for a literature translator or a second language learner. Here is a detailed example of the common and dependent properties of translation word pairs “山” and “mountain”. The two concepts share 8 common properties and differ in more properties as shown in table 5.

TABLE 5. The Cognitive Properties of “山-mountain” in Chinese and English with frequencies

CHS-Dependent	山 VS. mountain		ENG-Dependent
高 high-196	Common Properties		immovable-100
高耸 high-149	CHS	ENG	dignified-4
深重 deep&heavy-85	沉重-153	heavy-7	determined-3
多 many-50	重-37	heavy-7	hyaloid-3
高大 high-27	稳重-34	heavy-7	insensate-2
执着 persistence-26	大-31	big-2	bottleful-2
平静 calm-9	沉稳-24	heavy-7	earthbound-1
坚实 stable-9	坚定-8	staunch-1	foggy-1
挺拔 upright-9	伟岸-7	stalwart-1	phrasal-1
坚忍不拔 fortitudinous-8	坚强-6	staunch1	nonliving-1
崇高 sublimity-6			converse-1
…			…

In English, the most important property of mountain is “immovable” while it is “high” in Chinese. [1] The contrast is very useful in cross language teaching and communications. The automatic comparison is not very precise yet, we need to enlarge the scale of the cognitive property knowledgebase.

5 Conclusion and Future Work

Cognitive properties of words are very meaningful and useful but are not given in the traditional dictionaries. To overcome the difficulty in manual collecting, tagging and comparing of the cognitive properties in different languages, we employ search engines and bilingual dictionaries to construct an English-Chinese cognitive property knowledgebank. With the frequencies of the “vehicle-adjective” items, it is fast and convenient to see the language common and dependent properties of the word-pairs, which have translation relations. Using HowNet, we’ve already seen that most of the “vehicle-adjective” items are language dependent. Thus, the knowledgebank is very helpful to literature translators, language learners and machine translations.

In the future, we are to find better ways to collect more “vehicle-adjective” items from search engines and to use larger bilingual dictionaries to refine the common parts of English and Chinese cognitive properties. With more multi-lingual dictionaries, we are also able to deal with more languages under different cultures.

[1] The item “mountain-high” does not exist in our collection but appears in Google. Because it is hard to get the item only using the template “as adjective as”.

Acknowledgments

We are grateful for the comments of the anonymous reviewers. This work was supported in part by National Social Science Fund of China under contract 10CYY021, 11CYY030, State Key Lab. for Novel Software Technology under contract KFKT2011B03, China PostDoc Fund under contract 2012M510178, Jiangsu PostDoc Fund under contract 1101065C.

References

Dong, Z. D. & Dong, Q. 2006. *HowNet and the Computaion of Meaning*. Singapore, World Scientific Press,.

Miller, G. A., R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. WordNet: An online lexical database. *Int. J. Lexicograph.* 3, 4:235-244.

Jia Y. X. and Yu S. W. 2009. Instance-based Metaphor Comprehension and Generation. *Computer Science*, 36(3):138-41.

Kintsch, W. 2000. Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, volumn7: 257-66.

Lakoff, G. & Johnson. M. 1980. *Metaphors We Live by*. Chicago: The University of Chicago Press.

Roncero, C., Kennedy, J. M., and Smyth, R. 2006. Similes on the internet have explanations. *Psychonomic Bulletin and Review*, 13(1).

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