Using Similes to Extract Basic Sentiments across Languages

Bin Li[1,2], Haibo Kuang[2], Yingjie Zhang[1], Jiajun Chen[1], and Xuri Tang[3]

[1] State Key Lab for Novel Software Technology, Nanjing University,
210046 Nanjing, P.R. China
{lib,zhangyj,chenjj}@nlp.nju.edu.cn
[2] Research Center of Language and Informatics, Nanjing Normal University,
210007 Nanjing, P.R. China
mchypocn@hotmail.com
[3] Foreign Language School, Huazhong University of Science and Technology,
430074 Wuhan, P.R. China
xrtang@126.com

Abstract. People often use similes of pattern "as adjective as noun" to express their feelings on web medias. The adjective in the pattern is generally the salient property and strong impression of the noun entity in the speaker's mind. By querying the simile templates from search engines, we construct a large database of "noun-adjective" items in English and Chinese, which highlight the same and different basic sentiments on the same entity in the two languages. The approach is a fast and efficient way to extract people's basic sentiments and feelings across languages.

Keywords: sentiment extraction, simile, natural language processing.

1 Introduction

Sentiment extraction is a heated topic in natural language processing. People's attitudes towards a person, product, organization and other objects are potentially of great value to business intelligence and decision making systems[1]. Most of the previous researches focused on the supervised or unsupervised machine learning algorithms for the constructions of monolingual or multilingual lexical resources, corpus and systems[2]. However, there is one important and simple phenomenon that is not fully explored but able to provide new access to the knowledge. People would use the salient sentiments of an object in simile expressions, which could be used to detect people's sentiments. As shown in example(1,2,3), the three real sentences express strong attitudes on *BlackBerry apps*, *Mike* and *he*, which are the key topics to extract in sentiment research. But we want to emphasize that, the *iPhone*, *Kobe* and *pig* are more basic and important topics with their salient properties *polished*, *selfish* and *lazy*. These properties are the speaker's sentiments on the nouns.

(1) BlackBerry apps tend not to look as ***polished*** as ***iPhone***.
(2) Mike was just as ***selfish*** as ***Kobe***.
(3) He is as ***lazy*** as a ***pig***.

F.L. Wang et al. (Eds.): WISM 2012, LNCS 7529, pp. 536–542, 2012.

It is obvious that, the adjective used in a simile template "X be as ADJECTIVE as VEHICLE" precisely describes the property of the vehicle. In linguistics, the template can serve as comparison (example 1, 2) or simile(example 3) or irony(as fast as a snail), but the vehicle has the property expressed by the adjective in both the first two cases. So the simile templates are very useful in acquiring the vehicle nouns' properties. If tens of thousands of simile expressions are collected, a great number of vehicle-adjective items will be gathered. Then it is easy to obtain the different sentiments and properties of the same vehicle like *iPhone*-polished, coveted, fancy, etc.

The similes do not often occur in news and technique texts but in web blogs, micro blogs and forums. Thus, using search engine is a good way to get the similes fast and efficient. For the purpose of extracting basic attitudes cross languages, we employ simile sentences to collect "vehicle-adjective" items in English and Chinese. In this way, we construct a large database which reveals the shared and dependent attitudes on entities in the two languages.

2 Related Work

Cross language sentiment extraction has been experimented on parallel texts[3], translated texts[4] and comparable corpus[5]. All these work employed machine learning methods to classify or extract the sentiments in the corpus of comments, reviews and blogs.

Turney put forward a fast method to obtain the polarity of a word by querying the word from search engine with good and bad words[6]. Point-wise mutual information(PMI) is calculated as the polarity score. But it cannot supply the detailed attitudes of people on the entity word.

To collect similes from search engines is the interest of the metaphor researchers. Veale collects English "noun-adjective" pairs by querying the simile templates "as adjective as *" and "as * as noun" from Google(google.com) with WordNet to construct the English lexical metaphor knowledgebase "sardonicus", which contains about 10,000 items of "noun-adjective" pairs[7]. Similarly, Jia collects 20,000 items of Chinese "noun-adjective" pairs with similes from Baidu(Baidu.com)[8]. Thus, querying search engine is an efficient way to collect "noun-adjective" pairs. However, they concentrate on the adjectives of the common nouns in WordNet-like dictionaries, but not the entity words like "iPhone" and "Kobe". Second, they separately do the collecting work in one language. Third, they don't pay enough attention the frequencies of the items. Therefore, we want to extend the research to multi-languages and further the research with stress on the frequencies to show the sentiments distributions.

3 Simile Extraction

As the methods applied in [7] and [8], we use the specific simile templates to collect English and Chinese "noun-adjective" pairs by querying the search engines. The words in WordNet[9] and HowNet[10] are used for querying the search engines.

3.1 English Simile Extraction

We use the 21,479 adjectives in WordNet to fill in the simile template "as ADJ as". Google advanced search is queried with 3 limitations to refine the search results: exact search, English language and up to 100 results for each query. Then, 585,300 "as…as…" items(1,054,982 tokens) are obtained, many of which are nonsense, noisy and error items. We simply use the adjectives in HowNet as the filter to trim these items. Functional words, pronouns and the vehicles longer than 2 words are trimmed off. Only 98,057 types(178,622 tokens) of "noun-adjective" pairs are left, covering 12,468 adjectives in WordNet and 68,752 vehicles. Table 1 gives the top 10 most frequent pairs with their frequencies.

Table 1. Top 10 most frequent vehicle-adjective pairs in English

Id	Vehicle	Adjective	Freq
1	blood	red	628
2	twilight	gay	466
3	grass	perennial	413
4	ice	cold	392
5	mustard	keen	385
6	**Barack Obama**	**Irish**	358
7	snow	white	340
8	sea	boundless	314
9	feather	light	289
10	night	black	280

The "blood-red" is the most frequent item in English. And "Barack Obama-Irish" gets a high rank in the results. The frequency tells the salience of people's feelings of the vehicles. "Barack Obama" also have ten other adjectives like inexperienced(4 times), socialistic(once), etc. The short name "Obama" gets many more results like liberal(26 times), funny(14 times), and black(6 times), etc. These items are very useful as they take the basic and strong impressions in people's minds.

The frequencies here are not the exact data on the web. They are only the statistical situation in the collected items. And the frequency of the item can be over 100, because the item also occurs in the querying results of other words.

3.2 Chinese Simile Extraction

For Chinese, 3 simile templates "像(as)+NOUN+一样(same)", "像(as)+VERB+一样(same)", "像(as)+一样(same)+ADJ" are filled with the 51,020 nouns, 27,901 verbs and 12,252 adjectives in HowNet to query the Chinese search engine Baidu(baidu.com). Verbs are considered as vehicles, because Chinese words do not have inflections and some of verbs may serve as a noun in some context. We submit 91,173 queries to Baidu advanced search, setting that up to 100 results returned for

each query. As a result, 1,258,430 types(5,637,500 tokens) of "vehicle-adjective" pairs are gathered. Then, adjectives in HowNet are used to filter these items, leaving only 75,336 items. The 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. Top 10 most frequent vehicle-adjective pairs in Chinese.

Id	Vehicle	Adjective	Freq
1	苹果apple	时尚fashionable	1445
2	宝钗lady name	懂事reasonable	998
3	可卿lady name	漂亮pretty	943
4	美玉fine jade	美丽beautiful	840
5	呼吸breath	自然natural	758
6	晨曦sun rise	朝气蓬勃spirited	750
7	纸paper	薄thin	660
8	雨点rain drop	密集dense	557
9	自由freedom	美丽beautiful	543
10	雪snow	白white	521

It is surprising to see that the products of "apple (Inc.)" have taken the first place in Chinese eyes on the web media. And the two ladies named "宝钗" and "可卿" in the famous Chinese novel "A Dream of Red Mansions" get the second and third place. The rest of the vehicles in top10 items are common nouns. The rest of the adjectives of "apple" are 红(red,68 times) and 可爱(lovely, 25times), etc. Most of them do not refer to the company but the fruit.

In next section, we will compare the basic sentiments based on the collected data from Google and Baidu.

4 Bilingual Comparison of Basic Sentiments

In the last 2 sections, we've got the basic sentiments in English and Chinese. Now, it is natural to see if the sentiments are the same between them. However, we still lack a large dictionary or ontology of entity words and the sense tagging tool for disambiguation the senses of the vehicles. So what we can supply is to search these named entity words in the database bilingually with Google (translate.google.cn) or Bing (dict.bing.com.cn) translation service.

We randomly select 13 words of famous persons, products and companies to see their sentiments in people's eyes bilingually. The nouns in table 3 are common topics in English, including 3 famous persons, 3 products of apple and 7 enterprises. Most of the properties are correct, with few polysemous words like "Kobe-marbled"[1] and "apple-round". The bold adjectives are the same property shared by English and Chinese speakers.

[1] "Kobe" is also the name of a city in Japan.

Table 3. Bilingual comparison of sentiments on 13 entities

Noun_# of adjs	Top8 adjs with frequencies and Chinese translations
Obama_35 奥巴马_5	liberal_26,funny_14,magnanimous_6,black_6,ineligible_4,incompetent_4,phony_4,fake_4
	横 peremptory_15,单纯 pure minded_5,健美 good look_3,**好 good_2**,优秀 excellent_1
Messi_6 梅西_7	talented_8,acceptable_4,midget_2,reliant_5,skilful_2
	伟大 great_16,灵活 nimble_4,过人 outstanding_3,致命 deadly _2,低调 low-key _2,好 good_1,灵便 nimble_1,高 high level_1
Kobe_7 科比_8	cocky_6,[marbled_6],arrogant_2,**selfish_2**,disliked_1,emphatic_1,streaked_1
	强 strong_11, 坚强 strong heart_5,成功 successful_3, 准 accurate_2,出色 outstanding_1,孤独 lonely_1,勇猛 brave_1,自私 **selfish_1**
iPod_9 iPod_8	portable_4,**common_2**,swarthy_1,preferred_1,noteworthy_1,intuitive_1,important_1,dinky_1
	流行 popular_10,多彩 colorful_7,容易划伤 delicate _4,简单 simple_2,豪华 luxury_2,完胜 success_1,容易使用 easy to use_1,火 heated_1
iPad_24 iPad_4	lambent_8,diverse_5,slippery_5,intuitive_4,less_4,fast_3,nascent_2,telepathic_2
	流行 popular_5,持久 long haul_4,精致 exquisite_2,轻松 easy_1
iPhone_80 iPhone_8	useless_18,polished_6,mature_5,responsive_5,coveted_4,modal_4,popular_4,insecure_4,smooth_4,discreet_3,
	方便 convenient_2,成功 successful_2,供不应求 short supply_1,薄 thin_1,坚挺 strong_1,帅 cute_1,拉风 cool_1, 炫 dazzle_1
apple_62 苹果_30	**round_49**,modern_12,cool_6,light_6,**loved_6**,**pretty_6**,closed_5,**crisp_4**
	时尚 fashionable_1445,红 red_68,**可爱 lovely_25**,甜 sweet_20,圆 round_17,坚实 firm_12,**鲜嫩 fresh_8**,鲜红 red_7
Microsoft_22 微软_4	**sneaky_16**,commercialized_4,rugged_4,arrogant_2,inadequate_2,evil_2,ephemeral_2, monopolistic_2
	过时 outdated_3,成功 successful_2,**强大 powerful_1**,**无耻 sneaky_1**
Google_28 谷歌_6	mighty_16,permanent_4,acquisitive_3,omniscient_2,relevant_2,international_2,intelligent_2,helpful_2
	简洁 simple_19,**家喻户晓 well known_3**,年轻有为 bright young_1,强大 powerful_1,好 good_1,财大气粗 rich_1
Facebook_56 Facebook_2	social_40,bitchy_6,boring_6,creepy_6,indispensable_6,malign_6,erudite_4,extraordinary_4
	成功 successful_2,好 good_1
KFC_4 肯德基_2	sly_4,fried_2,fortified_1,synergistic_1,
	好吃 delicious_2,恶心 sick_1
McDonalds16 麦当劳_5	fattening_10,inhumane_6,mammoth_2,patronized_2,prevalent_2,caloric_2,corporate_2,dying_2
	方便 convenient_8,强大 powerful_8,正常 nomal_6,火 heated_3,好吃 delicious_1
Walmart_7 沃尔玛_3	classy_4,scummy_2,sanctified_2,predatory_1,exploitive_1,communist_1,avaricious_1
	大 large_5,多 many_3,实惠 boon_1

It is obvious that, most of the properties are different in the two languages. They are language or culture dependent, and the number of adjectives in English is almost larger than in Chinese. For the famous persons, people have positive and negative sentiments on them. *Obama* is somewhat welcomed in China, *Messi* is perfect without negative comments. *Kobe* is a great player, but his selfish is criticized cross languages. For apple's products, people love them very much. To have an apple product is a fashion in China, while the west users are not very satisfied with it. The three great companies *Apple*, *Google* and *Facebook* are enjoyed by the 2 language users, while *Microsoft* is getting worse. The two fast food companies and supermarket company *Walmalt* are still welcomed in China, but get more criticisms in the west.

The evaluation of the sentiment data is not easy to conduct. First, it is almost impossible to score the recall rate of the "vehicle-adjective" items, because we cannot know exactly how many adjectives people will use to describe the objects. Second, due to the large scale of the data, we can only manually check some samples to see the accuracy of the items. And the accuracy is high on the randomly selected 13 nouns. In the future, we need to design better methods for the evaluation.

5 Conclusion and Future Work

Extracting multi-lingual sentiments from web is a useful but hard task in natural language processing, because it has to face the efficiency and accuracy in lexical analysis, parsing, word sense disambiguation and machine translation. To avoid the complexity and low efficiency of the traditional methods, we put forward an easy and fast way to extract basic sentiments from search engines. Simile templates are accurate to obtain the adjectives expressing people's feelings and search engines are easy to query. Using Google and Baidu, we collected a database of tens of thousands "vehicle-adjective" pairs and then conducted filtering by phrase length and adjectives. The database constructed in the process supply the bilingual basic sentiments of a given topic.

In the future, we will collect bilingual entity dictionaries and ontologies to make full analysis of the basic sentiment database and supply online search service. We also want to expand our method to collect basic sentiments in more languages. At that time, it will become easy to browse the basic attitudes of things all over the world.

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, China PostDoc Fund under contract 2012M510178, State Key Lab. for Novel Software Technology under contract KFKT2011B03, Jiangsu PostDoc Fund under contract 1101065C.

References

1. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)

2. Liu, B.: Sentiment Analysis and Subjectivity. In: Handbook of Natural Language Processing, 2nd edn., Chapman and Hall/CRC (2010)
3. Ni, X., Sun, J.-T., Hu, J., Chen, Z.: Mining Multilingual Topics from Wikipedia. In: Proceedings of the 18th International Conference on World Wide Web, New York, USA (2009)
4. Wan, X.: Co-training for Cross-lingual Sentiment Classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Stroudsburg, PA, USA, vol. 1 (2009)
5. Boyd-Graber, J., Resnik, P.: Holistic Sentiment Analysis across Languages: Multilingual Supervised Latent Dirichlet Allocation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 45–55 (2010)
6. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the ACL, pp. 417–424 (2002)
7. Veale, T., Hao, Y.F.: Learning to Understand Figurative Language: From Similes to Metaphors to Irony. In: Proceedings of CogSci 2007, Nashville, USA (2007)
8. Jia, Y.X., Yu, S.W.: Instance-based Metaphor Comprehension and Generation. Computer Science 36(3), 138–141 (2009)
9. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: An Online Lexical Database. Int. J. Lexicography 3(4), 235–244 (1990)
10. Dong, Z.D., Dong, Q.: HowNet and the Computation of Meaning. World Scientific Press, Singapore (2006)