Chinese Emotion Lexicon Developing via Multi-lingual Lexical Resources Integration

Jun Xu[1], Ruifeng Xu[1,*], Yanzhen Zheng[1], Qin Lu[2],
Kai-Fai Wong[3,4], and Xiaolong Wang[1]

[1] Key Laboratory of Network Oriented Intelligent Computation,
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3] Department of SEEM, The Chinese University of Hong Kong, Hong Kong
[4] Key Laboratory of High Confidence Software Technologies, Ministry of Education, China
{xujun,xuruifeng}.hitsz.edu.cn, csluqin@comp.polyu.edu.hk,
kfwong@se.cuhk.edu.cn, xlwang@insun.hit.edu.cn

Abstract. This paper proposes an automatic approach to build Chinese emotion lexicon based on WordNet-Affect which is a widely-used English emotion lexicon resource developed on WordNet. The approach consists of three steps, namely translation, filtering and extension. Initially, all English words in WordNet-Affect synsets are translated into Chinese words. Thereafter, with the help of Chinese synonyms dictionary (Tongyici Cilin), we build a bilingual undirected graph for each emotion category and propose a graph based algorithm to filter all non-emotion words introduced by translation procedure. Finally, the Chinese emotion lexicons are obtained by expanding their synonym words representing the similar emotion. The results show that the generated-lexicons is a reliable source for analyzing the emotions in Chinese text.

Keywords: Emotion lexicon development, Emotion analysis, Multi-lingual.

1 Introduction

Sentiment analysis studies how to identify and extract the subjective information in text. It may be divided into two aspects which are opinion and emotion, respectively. Opinion, in general, is the judgment or evaluation of a speaker or a writer with respect to some topic, such as negative/positive, pros/cons. While emotion is a strong human feeling, the emotional state of a person, such as joy, anger, sadness, fear, etc.

With the popularity of the social network, social media plays an important role in the information release and dissemination. It has been a new and a good media platform for fast and wide spread of information nowadays. To measure and recognize the emotional changes of population in large scale is one of the most important areas in social sciences [1] and economics studies [2]. Therefore, text emotion analysis research attracts much attention. Many emotion analysis approaches essentially rely

* Corresponding author.

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 174–182, 2013.

on emotion lexical resources containing words and their associated emotions. Thus, the establishing of emotion lexicon is recognized as the foundation in the research of emotion analysis.

However, the emotion lexicon is still not easily available for Chinese or other resource poor languages. This paper focuses on the automatic construction of a Chinese emotion lexicon, starting from WordNet-Affect which is a widely used emotion lexicon in English, through the translation and integration with multi-lingual lexical resources.

The rest of this paper is organized as follows. Section 2 gives a brief review on the construction of emotion lexical resource. Section 3 presents a brief description for two lexical resources which will be used in this study. Section 4 introduces our proposed three steps automatic Chinese emotion lexicon construction approach. Section 5 gives the performance evaluation and Section 6 concludes.

2 Emotion Lexicon Construction: State of the Art

The emotion lexicons can be used as semantic knowledge base for emotion analysis. For the development of emotion lexicons, there are two questions faced: which word can be used to express emotions? and what kind of emotion or set of emotions that the words convey? In the previous studies on emotion lexicon construction, there are two major approaches adopted: extension from semantic lexical resource and corpus-based extraction with heuristic rule.

2.1 Extension with Semantic Lexical Resource

To create an emotion lexicon automatically, the existing lexical resource may be a good starting point. Starting with WordNet, Strapparava and Valitutti developed WordNet-Affect [3].Several seed emotion words are manually chose and then the correlation of relations defined in WordNet (e.g., causes, entailment and so on), emotional tags and domain tags are used to expand. Finally, WordNet-Affect, the collection of emotion synsets are obtained by exploiting associated affection. With WordNet-Affect, many researchers attempt to expand it to other languages for developing multi-lingual emotion lexicons. Sokolova and Bobicev [4] translated every word in WordNet-Affect to Romanian and Russian. They used three machine learning methods to classify the emotions of these words which are represented by the word spelling and word form. Torii et al. [5] constructed a Japanese WordNet-Affect directly according to WordNet-Affect's SyssetID by making use of Japanese WordNet.

2.2 Corpus-Based Extraction with Heuristic Rule

Using emoticons (such as ":)" and ": o") in the blogs as the clues, Yang et al. [6] exploited co-occurrence based algorithm in collocations to extract emotion words from blog corpora. Xu L. et al. [7] built a Chinese affective lexical ontology. The emotions of the ontology are hierarchical categorized into 7 categories on first level and 20 categories on second level. They annotated the emotion label and intensity for each emotion word, which are manually collected from related semantic lexicons. In

contrast to the method of expanding from semantic lexicon, they finally labeled the emotion category and compute the intensity for all candidate words automatically based on mutual information on a large corpus. Xu G. et al. [8] proposed a graph-based approach to identify the emotion label of a word. They computed the similarity between the candidate words and seed words with different similarity metrics by leveraging un-annotated corpora, lexicon resources, heuristic rules and so on. Thereafter, they built the word similarity matrices after integration to label each candidate word iteratively based on their proposed graph-based algorithm. Quan and Ren [10] identified the emotion words by training a Maximum Entropy based classifier on an emotional labeling corpus, Ren-CECps [9] with semantic features.

3 Lexical Resources Used in this Research

3.1 English Emotion Lexicon – WordNet-Affect

The English WordNet-Affect is a widely used emotion lexical resource with affective annotation. It was developed on WordNet based on Ekman's six emotion types (anger, disgust, fear, joy, sadness, surprise) theory. WordNet-Affect is a subset of WordNet which contains the essential knowledge related to emotion analysis.

WordNet-Affect is provided in six files named by the six emotions, respectively. Each file lists the synsets they contain per line. Following line is an example synset entry in WordNet-Affect.

```
n#05588822 umbrage offense
```

In this line, the first letter gives the part of speech (POS) of this entry and it is followed by the synset ID, and then the synonyms in this synset.

3.2 Chinese Synonym Dictionary – Tonngyici Cilin

Tongyici Cilin (in short Cilin) is a well-used Chinese synonyms dictionary, which was published in 1983. It contains about 50 thousands of Chinese words. Three-level conceptual categories are adopted to cluster synonyms according to their semantic relationships. The top level category consists of 12 main classes. The second level category consists of 94 classes. While the third level concepts are classified into 1428 classes.

In this study, we use HIT IR-Lab Tongyici Cilin[1] (Extended)(EClin for short) as Chinese lexical resources. ECilin extended the three-level categories of original Tongyici Cilin to five levels while the rare and unusual words are filtered out. At the same time, some new words are added in. In ECilin, a capital letter is used to label the fourth level, which is the concept clusters. The deepest one stands for atomic concepts, in which words are nearly synonyms. There are three tags: "=" stands for the same sense; "#" for antonyms and "@" for enclosure which means the word has no synonyms. An example entry of ECilin is given below. All words "欢腾 欢跃 手舞

[1] http://ir.hit.edu.cn

足蹈 欢呼雀跃" (clam happy) of the entry are followed by a sense code "Ga01A04", the five-level categories. The words having the same sense code can be regarded as of similar meaning.

Ga01A04=欢腾 欢跃 手舞足蹈 欢呼雀跃

4 Our Approach

In this section, our approach for constructing a Chinese emotion lexicon is presented. This approach contains three steps: translation, filtering and extension.

4.1 Translation

The goal of translation is to translate English emotion words in WordNet-Affect to Chinese as the emotion word candidates as much as possible for the following procedure. To translate each word in the WordNet-Affect synsets from English to Chinese, two online machine translation systems are used, i.e., BaiDu Translator[2] and YouDao Translator[3]. Both of them are well-known and widely used in Chinese-English translation area, as well as they support free API. The YouDao Translator outputs all translations with corresponding part-of-speech (POS) tags. In this study, the target translated words whose POS match the source word's POS are returned for following procedure.

Table 1. An example of WordNet-Affect synsets translation

WordNet-Affect Synset	n#05588321 wrath	n#05588822 umbrage	n#05588822 offense
A, Translated Results from BaiDu	NULL	阴影,树荫;簇叶,愤怒;生气	罪过;犯法;过错冒犯;触怒,引起反感的事物
B, Translated Results from YouDao	愤怒; 激怒	不快;生气,树荫,怀疑	犯罪;过错,进攻,触怒,引起反感的事物
A ∪ B	愤怒; 激怒	阴影;树荫;簇叶;愤怒;生气;不快;怀疑	冒犯;引起反感的事物;犯法;犯罪;罪过;触怒;过错;进攻

Table 1 demonstrates an example of the synset translation procedure. In Table 1, "NULL" denotes that there is no returned translations since there are some words in the English synset can't be translated to Chinese words. To ensure the integrity of translation procedure, the union of all outputs from different machine translation systems, i.e., A ∪ B is admitted.

[2] http://fanyi.baidu.com/
[3] http://fanyi.youdao.com

4.2 Filtering

The original English words in a synset of WordNet-Affect have similar meanings, while their corresponding Chinese translations have much ambiguity as there is no way to obtain accurate equivalent words during the translation. In the translation step, all possible translations for all of their senses are provided. It means that many noisy or irrelevant words are introduced. For instance, the Chinese words "树荫(shade of tree)", "簇叶(foliage)" in the translations of synset "n#05588822". Such kind of words should be filtered. Therefore, we propose a bilingual undirected graph based filtering algorithm for automatic sense disambiguation. The Chinese words in the translated synsets which convey the same emotion will be figured out.

For each emotion category, the bilingual graph G is constructed as follows:

Step 1 Create a R-vertex graph with no edges. R is a starting vertex.
Step 2 Let $S = \{s_1, s_2 \ldots s_n\}$ denotes the set of synsets, s_i is synset ID. For each $s_i \in S$, add s_i as a synset vertex and add an edge (R, s_i) between R and a synset s_i.
Step 3 Let $E = \{e_1, e_2 \ldots e_m\}$ denotes the synonyms in S, For each $e_j \in E$, add e_j as an English word vertex and add an edge (e_j, s_i) if and only if e_j belongs to synset s_i.
Step 4 Lets $C = \{c_1, c_2 \ldots c_l\}$ denotes all translated Chinese words, For each $c_k \in C$, add c_k as a Chinese word vertex and add an edge (c_k, e_j) if and only if c_k is in e_j 's translation results.
Step 5 For all Chinese word vertices, if two words are synonyms, then add an edge between them.

ECilin is utilized in this research for synonym judgment.

Figure 1 shows a partial bilingual graph after adding edges to link the synonyms. As shown in Figure 1, each simple path[4] between R and a Chinese word vertex include a synset and an English word vertex. If a Chinese word vertex has at least two simple paths to reach the vertex R, and these paths go through different synset and English vertex, the Chinese word can be treated as emotion word. It means that such Chinese word may share the same emotion sense in different English synsets. For example, "愤怒", "激怒", "不快" , "生气", "触怒", "冒犯" vertices in Figure 1 are classified as members of emotion lexicon of "anger" in our filtering algorithm.

The pseudo code of the proposed bilingual graph based filtering algorithm is given below. The Chinese word c is treated as a terminal vertex. We use depth first search to detect all simple paths between the start vertex R and c firstly (line 2). If there are at least two paths which do not have same synset and English word vertices, c can be annotated as an emotion word with a corresponding label. (line 3-5).

[4] A simple path is a path with no repeated vertices.

4.3 Extension

With the above procedures, six emotion lexicons are obtained corresponding to each emotion category, respectively. As shown in Table 2, the words for each emotion are very few in number. There are 220 unique words in "Anger" category, 58 words in "Disgust", 152 words in "Fear", 516 words in "Joy", 200 words in "Sadness", and 67 words in "Surprise". Obviously, current emotion lexicon is not efficient enough for a practical application of Chinese emotion analysis. Naturally, our aim is further extend current lexicon.

In this study, ECilin is utilized here for lexicon extension. For all words in current emotion lexicons, if it is found in ECiliin, all of the words with the same sense code are added to the corresponding emotion lexicon. Generally speaking, the words with the same sense code which have higher hit frequency have more relevance to the corresponding emotion, and thus they may be added to this emotion lexicon. After the extension procedure, the lexicon of each category has a great increment in number as Table 2 shows.

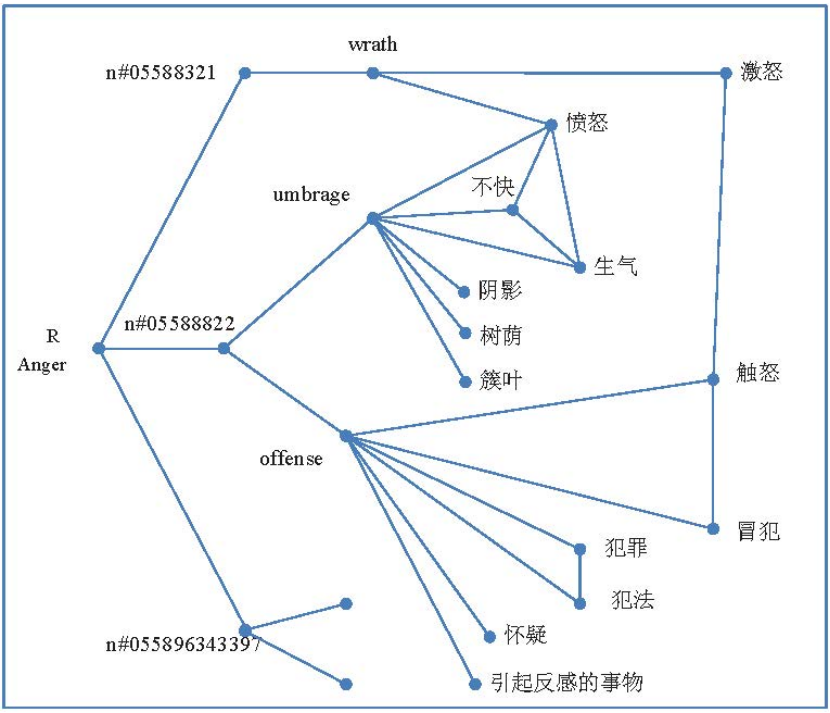

Fig. 1. A partial bilingual graph of "Anger"

Algorithm 1. The bilingual graph based filtering algorithm

```
Input: G, the constructed bilingual graph
       S, set of synset vertices
       E, set of English word vertices
       C, set of Chinese word vertices
Output: O, set of Chinese emotion words
1: for each word c ∈ C do
2:    Use the Depth First Search to find the paths set
         p = {p(R,c)|all simple paths between R and c};
3:    if ∃ p1 ∈ P and ∃ p2 ∈ P, while p1 ≠ p2, p1 ∩ p2 ∩ S = ∅
         and p1 ∩ p2 ∩ E = ∅
      then
4:          Add c to O;
5:    end if
6:  End for
```

5 Evaluation and Analysis

As we know, the emotion carried by a word is inherently uncertain and subjective. To evaluate the quality of the obtained emotion lexicons, manual judgment are performed. Three raters annotate each automatically generated Chinese emotion word independently. After annotations, we estimate the pairwise kappa of emotion tags among them to evaluate the tagging quality. For generation the final lexicons, we use a lenient standard. If two of three raters have same annotation, the word is accepted. The final lexicons after agreement serve as the gold standard. Precision (P) is adopted as the evaluation metric and it is computed as

$$\frac{\#human\_corrected}{\#system\_propsed} \times 100\%$$

Table 2. All results of proposed approach

Emotion Category	After Translation (Num. of Words)	After Filtering (Num. of Words)	After Extension (Num. of Words)	After Agreement (Num. of Words)	P
Anger	525	220	1022	852	0.8337
Disgust	144	58	1084	926	0.8542
Fear	354	152	493	380	0.7708
Joy	993	516	1838	1737	0.9450
Sadness	394	200	1493	1357	0.9089
Surprise	194	67	613	384	0.6264
Total	2604	1213	6543	5667	0.8614

Table 2 shows the number of words for each emotion category after agreement by raters as well as the precision. It is observed that our proposed approach achieves a

good precision. It is also observed that the lexicon generation for some specific emotion, such as “Surprise”, achieves lower precision. As we know, a word or phrase may express more than one emotion. For example, the idiom “惊慌失措 (dismayed)” expresses both “surprise” and “fear”. Statistics on Ren-CECps also shows that about 15.1% Chinese emotion words are multi-emotion ones which express complex feelings in its usage [10]. After agreement, 218 multi-emotions words are kept.

Table 3. Inter-rater agreements by category on generated lexicons

Emotion Category	Num. of Words	Averaged Kappa(K)
Anger	1022	0.6398
Disgust	1084	0.5772
Fear	493	0.6864
Joy	1838	0.3865
Sadness	1493	0.6399
Surprise	613	0.5475
Macro-averaged		0.5796

Table 3 shows the averaged value of the kappa coefficient for each emotion category, respectively. The values vary from 0.5475 to 0.6864 except for the lexicon of “Joy". Though agreements vary within categories, the macro-averaged kappa value is near to 0.6. This value is considered as good performance which indicates a good level of agreement. It also states that the final obtained emotion lexicon after agreement is reliable. For the “Joy” lexicon, the Kappa value is lowest (i.e., 0.3865). However, as shown in Table 2, the “Joy” lexicon has a high precision (0.9450). This happens when Kappa deviates from the normal distribution. It ignores the high inter-observer agreement of the annotation result.

Compared to work of Xu, G. et al. in [8], the proposed approach achieved a much larger lexicon with good precision. Compared to the work reported in [7], our approach saves human labor in great deal. Furthermore, the construction process of our proposed approach is easy to repeat.

6 Conclusion

In this paper, we presented an approach for developing Chinese emotion lexicon by using a English emotion lexicon and a Chinese thesaurus. This lexicon is developed starting from a English emotion lexicon, WordNet-Affect, through the translation, filtering and extension. We translated the WordNet-Affect synsets into Chinese, and afterwards integrated with another Chinese thesaurus, Tongyici Cilin to filter irrelevant words and also to expand it. The obtained Chinese emotion lexicon is freely available at http://icrc.hitsz.edu.cn/emotion_lexcions.rar.

In the future, we will continue to enrich this resource to make it useful in affective computing and emotion-based human inter-action applications.

Acknowledgement. This work is supported by the China Postdoctoral Science Foundation (No. 2011M500670), National Natural Science Foundation of China (No. 61203378 and No. 61272383), Shenzhen Foundational Research Funding (NO. JCYJ2012 0613152557576) and General Research Fund of Hong Kong (No. 417112).

References

1. Dodds, P.S., Danforth, C.M.: Measuring the Happiness of Large Scale Written Expression: Songs, Blogs, and Presidents. Joural of Happiness Studies 11(4), 441–456 (2009)
2. Bollen, J., Mao, H., Zeng, X.-J.: Twitter Mood Predicts the Stock Market. Journal of Computational Science 2(1), 1–8 (2011)
3. Strapparava, C., Valitutti, A.: WordNet-Affect: An Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1083–1086 (2004)
4. Sokolova, M., Bobicev, V.: Classification of Emotion Words in Russian and Romanian Languages. In: Proceedings of the International Conference RANLP 2009, pp. 416–420 (2009)
5. Torii, Y., Das, D., Bandyopadhyay, S., Okumura, M.: Developing Japanese Word-Net Affect for Analyzing Emotions. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 80–86 (2011)
6. Yang, C., Lin, K.H.-Y., Chen, H.-H.: Building Emotion Lexicon from Weblog Corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume: Proceedings of the Demo and Poster Sessions, pp. 133–136 (2007)
7. Xu, L., Lin, H., Pan, Y., Ren, H., Chen, J.: Constructing the Affective Lexicon Ontology. Journal of the China Society for Scientific and Technical Information 27(2), 180–185 (2008)
8. Xu, G., Meng, X., Wang, H.: Build Chinese Emotion Lexicons Using a Graph Based Algorithm and Multiple Resources. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1209–1217 (2010)
9. Quan, C., Ren, F.: Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1446–1454 (2009)
10. Quan, C., Ren, F.: An Exploration of Features for Recognizing Word Emotion. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 922–930 (2010)