Building Chinese Affective Resources in Valence-Arousal Dimensions

Liang-Chih Yu[1,3]**, Lung-Hao Lee**[4]**, Shuai Hao**[5]**, Jin Wang**[2,3,6]**, Yunchao He**[2,3,6]**,**
Jun Hu[5]**, K. Robert Lai**[2,3] and **Xuejie Zhang**[6]
[1]Department of Information Management, Yuan Ze University, Taiwan
[2]Department of Computer Science & Engineering, Yuan Ze University, Taiwan
[3]Innovation Center for Big Data and Digital Convergence Yuan Ze University, Taiwan
[4]Information Technology Center, National Taiwan Normal University**,** Taiwan
[5]School of Software, Nanchang University, Jiangxi, P.R. China
[6]School of Information Science and Engineering, Yunnan University, Yunnan, P.R. China
Contact: lcyu@saturn.yzu.edu.tw

Abstract

An increasing amount of research has recently focused on representing affective states as continuous numerical values on multiple dimensions, such as the valence-arousal (VA) space. Compared to the *categorical* approach that represents affective states as several classes (e.g., positive and negative), the *dimensional* approach can provide more fine-grained sentiment analysis. However, affective resources with valence-arousal ratings are still very rare, especially for the Chinese language. Therefore, this study builds 1) an affective lexicon called Chinese valence-arousal words (CVAW) containing 1,653 words, and 2) an affective corpus called Chinese valence-arousal text (CVAT) containing 2,009 sentences extracted from web texts. To improve the annotation quality, a corpus cleanup procedure is used to remove outlier ratings and improper texts. Experiments using CVAW words to predict the VA ratings of the CVAT corpus show results comparable to those obtained using English affective resources.

1 Introduction

Sentiment analysis has emerged as a leading technique to automatically identify affective information from texts (Pang and Lee, 2008; Calvo and D'Mello, 2010; Liu, 2012; Feldman, 2013). In sentiment analysis, affective states are generally represented using either categorical or dimensional approaches.

The categorical approach represents affective states as several discrete classes such as positive, neutral, negative, and Ekman's six basic emotions (e.g., anger, happiness, fear, sadness, disgust and surprise) (Ekman, 1992). Based on this representation, various practical applications have been developed such as aspect-based sentiment analysis (Schouten and Frasincar, 2016; Pontiki et al., 2015), Twitter sentiment analysis (Saif et al., 2013; Rosenthal et al., 2015), deceptive opinion spam detection (Li et al., 2014), and cross-lingual portability (Banea et al., 2013; Xu et al., 2015).

The dimensional approach represents affective states as continuous numerical values in multiple dimensions, such as valence-arousal (VA) space (Russell, 1980), as shown in Fig. 1. The valence represents the degree of pleasant and unpleasant (i.e., positive and negative) feelings, while the arousal represents the degree of excitement and calm. Based on this representation, any affective state can be represented as a point in the VA coordinate plane. For many application domains (e.g., product reviews, political stance detection, etc.), it can be useful to identify highly negative-arousing and highly positive-arousing texts because they are usually of interest to many users and should be given a higher priority. Dimensional sentiment

Proceedings of NAACL-HLT 2016, pages 540–545,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

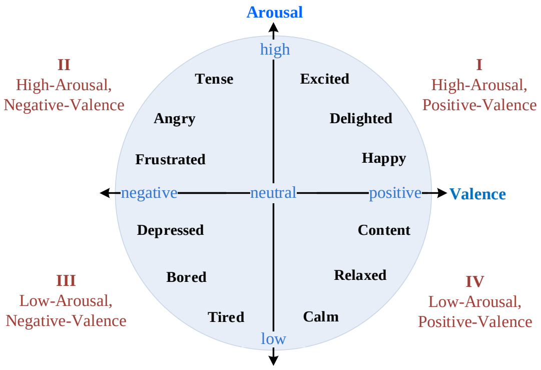

Figure 1: Two-dimensional valence-arousal space.

analysis can accomplish this by recognizing the valence-arousal ratings of texts and ranking them accordingly to provide more intelligent and fine-grained services.

In developing dimensional sentiment applications, affective lexicons and corpora with valence-arousal ratings are useful resources but few exist, especially for the Chinese language. Therefore, this study focuses on building Chinese valence-arousal resources, including an affective lexicon called the Chinese valence-arousal words (CVAW) and an affective corpus called the Chinese valence-arousal text (CVAT). The CVAW contains 1,653 affective words annotated with valence-arousal ratings by five annotators. The CVAT contains 2,009 sentences extracted from web texts annotated with crowd-sourced valence-arousal ratings. To further demonstrate the feasibility of the constructed resources, we conduct an experiment to predict the VA ratings of the CVAT corpus using CVAW words, and compare its performance to a similar evaluation of English affective resources.

To our best knowledge, only one previous study has manually created a small number (162) of Chinese VA words (Wei et al, 2011), and none have focused on creating Chinese VA corpora. This pilot study thus aims to build such resources to enrich the research and development of multi-lingual sentiment analysis in VA dimensions.

The rest of this paper is organized as follows. Section 2 introduces existing affective lexicons and corpora. Section 3 describes the process of building the Chinese affective resource. Section 4 presents the analysis results and feasibility evaluation. Conclusions are finally drawn in Section 5.

2 Related Work

Affective resources are usually obtained by either self-labeling or manual annotation. In the self-labeling approach, users proactively provide their feelings and opinions after browsing the web content. For example, users may read a news article and then offer comments. A user can also review the products available for sale in online stores. In the manual annotation method, trained annotators are asked to create affective annotations for specific language resources for research purposes. Several well-known affective resources are introduced as follows.

SentiWordNet is a lexical resource for opinion mining, which assigns to each synset of WordNet three sentiment ratings: positive, negative, and objective (Esuli and Sebastiani, 2006). Linguistic Inquiry and Word Count (LIWC) calculates the degree to which people use different categories of words across a broad range of texts (Pennebaker et al., 2007). In the LIWC 2007 version, the annotators were asked to note their emotions and thoughts about personally relevant topics. The Affective Norms for English Words (ANEW) provides 1,034 English words with ratings in the dimensions of pleasure, arousal and dominance (Bradley and Lang, 1999). In addition to these English-language sentiment lexicons, a few Chinese lexicons have been constructed. The Chinese LIWC (C-LIWC) dictionary is a Chinese translation of the LIWC with manual revisions to fit the practical characteristics of Chinese usages (Huang et al., 2012). The NTU Sentiment dictionary (NTUSD) has adopted a combination of manual and automatic methods to include positive and negative emotional words (Ku and Chen, 2007). Among the above affective lexicons, only ANEW is dimensional, providing real-valued scores for three dimensions, and the others are categorical, providing information related to sentiment polarity or intensity.

In addition to lexicon resources, several English-language affective corpora have been proposed, such as Movie Review Data (Pang et al. 2002), the MPQA Opinion Corpus (Wiebe et al., 2005), and Affective Norms for English Text (ANET) (Bradley and Lang, 2007). In addition, only ANET provides VA ratings. The above dimensional affective resources ANEW and ANET have been used for both word- and sentence-level VA prediction in previous studies (Wei et al., 2011; Gökçay et al., 2012; Malandrakis et al., 2013; Paltoglou et al., 2013; Yu et al., 2015). In this study, we follow the manual annotation approach to build a Chinese affective lexicon and corpus in the VA dimensions.

3 Affective Resource Construction

This section describes the process of building Chinese affective resources with valence-arousal ratings, including the CVAW and CAVT.

The CVAW is built on the Chinese affective lexicon C-LIWC, and then annotated with VA ratings for each word. Five annotators were trained to rate each word in the valence and arousal dimensions using the Self Assessment Manikin (SAM) model (Lang, 1980). The SAM model provides affective pictures, which can help annotators in determining more precise labels when rating the words. The valence dimension uses a nine degree scale. Values 1 and 9 respectively denote the most negative and positive degrees of affect. Point 5 means a neutral emotion without specific tendency. The arousal dimension uses a similar scale to denote calm and excitement Using this approach, each affective word can be annotated with VA ratings (determined by the average rating values provided by the annotators) to form the CVAW.

To build the CVAT, we first collected 720 web texts from six different categories: news articles, political discussion forums, car discussion forums, hotel reviews, book reviews, and laptop reviews. A total of 2,009 sentences containing the greatest number of affective words found in the C-LIWC lexicon were selected for VA rating. The Google app engine was then used to implement a crowdsourcing annotation platform using the SAM annotation scheme. Volunteer annotators were asked to rate individual sentences from 1 to 9 in terms of valence and arousal. Each sentence was rated by at least 10 annotations. Once the rating process was finished, a corpus cleanup procedure was performed to remove outlier ratings and improper sentences (e.g., those containing abusive or vulgar language). The outlier ratings were identified if they did not fall into the interval of the mean plus/minus 1.5 standard deviations. They were then excluded from the calculation of the average VA ratings for each sentence.

4 Results

4.1 Analysis Results of CVAW

A total of 1,653 words along with the annotated VA ratings were included in the CVAW lexicon, yielding the (mean, standard deviation) = (4.49, 1.81) for valence and (5.48, 1.26) for arousal. To analyze differences between the annotations, we compared the VA values rated by each annotator against their corresponding means across the five annotators to calculate the error rates using the following metrics.

- *Mean Absolute Error* (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | A_i - \overline{A_i} | ,$$

- *Root Mean Square Error* (RMSE):

$$RMSE = \sqrt{\sum_{i=1}^{n}\left(A_i - \overline{A_i}\right)^2 \Big/ n} ,$$

where A_i denotes the valence or arousal value of word *i* rated by an annotator, $\overline{A_i}$ denotes the mean valence or arousal of word *i* calculated over the five annotators, and *n* is the total number of words in the CVAW.

	MAE		RMSE	
	Valence	Arousal	Valence	Arousal
Annotator A	0.4934	1.3479	0.6372	1.6411
Annotator B	0.5972	0.7821	0.7488	0.9929
Annotator C	0.5817	1.1393	0.7423	1.4302
Annotator D	0.5188	0.8226	0.6614	1.0374
Annotator E	0.6258	1.0200	0.7970	1.2700
(Mean, SD)	(0.56,0.05)	(1.02, 0.21)	(0.72, 0.06)	(1.27, 0.24)

Table 1: Analysis of error rates of different annotators for building the Chinese VA lexicon.

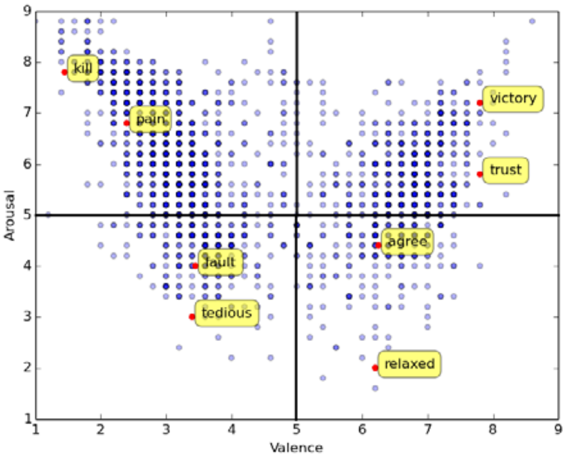

Figure 2: Scatter plot of the CVAW lexicon.

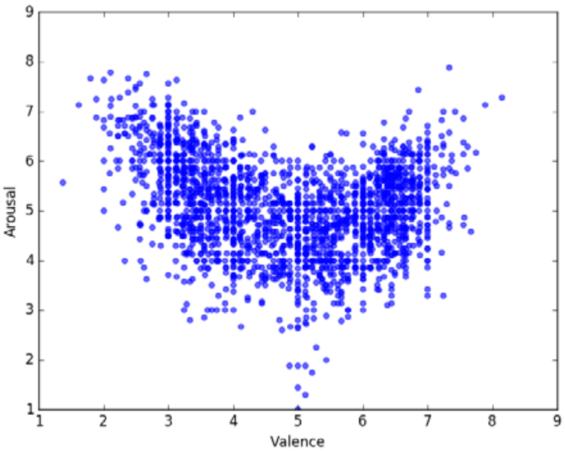

Figure 3: Scatter plot of the CVAT corpus.

	Num. of texts	Num. of tokens	Avg. tokens	Valence			Arousal		
				MAE	RMSE	*r*	MAE	RMSE	*r*
ANEW vs Forum	20	15,035	751.75	1.20	1.55	0.77	0.72	0.85	0.27
CVAW vs CVAT	2,009	70,456	35.07	1.20	1.52	0.54	1.01	1.28	0.16
Book Review	287(14%)	8,217	28.63	1.00	1.31	0.41	**0.89**	**1.11**	0.21
Car Forum	257 (13%)	12,261	47.71	1.48	1.77	0.30	0.92	1.15	0.10
Laptop Review	183 (9%)	5,374	29.37	**0.95**	**1.21**	0.55	1.07	1.40	0.04
Hotel Review	301 (15%)	7,268	24.15	1.35	1.73	0.59	0.93	1.17	**0.22**
News Article	542(27%)	21,923	40.45	1.11	1.40	**0.61**	1.11	1.40	0.17
Politics Forum	439 (22%)	15,413	35.11	1.28	1.61	0.51	1.04	1.32	0.19

Table 2: Results of using the CVAW lexicon to predict the VA ratings of the CVAT corpus.

Table 1 shows the error rates of the annotators in rating the VA values of words in the CVAW. Overall, for all metrics the error rates of arousal ratings were greater than those of valence ratings. In addition, the annotators produced more consistent error rates (around 0.49~0.63 for MAE and 0.64~0.80 for RMSE) in the valence dimension than those (around 0.78~1.35 for MAE and 0.99~1.64 for RMSE) in the arousal dimension. These findings indicate that the degree of arousal was more difficult to distinguish than valence.

Figure 2 shows a scatter plot of words in the CVAW, where each point represents the mean of the VA values as rated by the annotators. Several words (translated from Chinese) were marked in the VA space for reference, e.g., *victory* (7.8, 7.2), *trust* (7.8, 5.8), *pain* (2.4, 6.8), *kill* (1.6, 7.8), *tedious* (3.4, 3), *fault* (3.6, 4.6), *agree* (6.4, 4.4) and *relaxed* (6.2, 2.0).

4.2 Analysis Results of CVAT

A total of 2,009 sentences with VA ratings were included in the CVAT corpus, yielding the (mean, standard deviation) = (4.83, 1.37) for valence and (5.05, 0.95) for arousal. The distribution of the six categories and their word counts in CVAT are shown in Table 2. The largest category was News (27%), while the smallest one was Laptop (9%). Figure 3 shows a scatter plot of VA ratings for all sentences in CVAT. It is similar with the plot of the CVAW, indicating that annotators followed similar guidelines for rating affective words and sentences.

4.3 Results of Using CVAW to Predict the VA Ratings of CVAT

To demonstrate the application of the constructed affective resources, this experiment adopted a simple aggregate-and-average method (Taboada et al. 2011) to predict the VA ratings of the CVAT corpus using CVAW words. In this approach, the valence (or arousal) rating of a given sentence was calculated by averaging the valence (or arousal) ratings of the words matched in the CVAW in that sentence. Once the predicted values of the VA ratings for the sentences were obtained, they were compared to the corresponding actual values in the CVAT to calculate MAE, RMSE and Pearson correlation coefficient *r*, as shown in Table 2. Notice that the sentences which contain no affective words in the CVAW were not included for performance calculation (herein 30 sentences). The results using ANEW to predict the VA rating of 20 English forum discussions were also included for comparison (Paltoglou et al., 2013).

The results show that the average tokens of the CVAT sentences are around 35 which is much smaller than those of the English forum discussions (long texts). Both English and Chinese resources had a similar error rates (MAE and RMSE) for valence, while the English resource outperformed the Chinese resource in terms of arousal rates. In addition, both the English and Chinese resources had a lower correlation for arousal than for valence, indicating again that the arousal dimension is more difficult to predict. Table 2 also shows the performance for each category in CVAT. For valence, Laptop achieved the lowest error rate, while News and Hotel had a higher correlation. The respective ranges of MAE, RMSE and *r* are 0.95~1.48, 1.21~1.77 and 0.30~0.61. For arousal, Book yielded the lowest error rate, while Hotel and Book yielded a better correlation. The respective ranges of MAE, RMSE and *r* are 0.89~1.11, 1.11~1.40 and 0.04~0.22.

5 Conclusions and Future Work

This study presents a Chinese affective lexicon with 1,653 words and a corpus of 2,009 sentences with six different categories, both annotated with valence-arousal values. A corpus cleanup procedure was used to remove outlier ratings and improper texts to improve quality. Experimental results provided a feasibility evaluation and baseline performance for VA prediction using the constructed resources. Future work will focus on building useful dimensional sentiment applications based on the constructed resources.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan, ROC, under Grant No. NSC102-2221-E-155-029-MY3. The authors would like to thank the anonymous reviewers and the area chairs for their constructive comments.

References

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2013. Porting multilingual subjectivity resources across languages. *IEEE Trans. Affective Computing*, 4(2):211-225.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, University of Florida, Gainesville, FL.

Margaret M. Bradley and Peter J. Lang. 2007. Affective Norms for English Text (ANET): Affective ratings of text and instruction manual. Technical Report D-1, University of Florida, Gainesville, FL.

R. A. Calvo and Sidney. D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affective Computing*, 1(1): 18-37.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6:169-200.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: a publicly available lexical resource for opinion mining. In *Proc. of LREC-06*, pages 417-422.

Ronen Feldman. 2013. Techniques and applications for sentiment a0nalysis. *Communications of the ACM*, 56(4):82-89.

Didem Gökçay, Erdinç İşbilir and Gülsen Yıldırım. 2012. Predicting the sentiment in sentences based on words: an exploratory study on ANEW and ANET. In *Proc. of CogInfoCom-12*, pages 715-718.

C.-L. Huang, C. K. Chung, N. Hui, Y.-C. Lin, Y.-T. Seih, W.-C. Chen, B. Lam, M. Bond, and James W. Pennebaker. 2012. The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese Journal of Psychology*, 54(2):185-201.

Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the web: beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12), 1838-1850.

Peter J. Lang. 1980. Behavioral treatment and biobehavioral assessment: Computer applications. *Tech-*

nology in Mental Health Care Delivery Systems, pp. 119-137, Ablex Publishing, Norwood.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proc. of ACL-14*, pages 1566-1576.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, Chicago, IL.

Nikos Malandrakis, Alexandros Potamianos, Elias Iosif, Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Trans. Audio, Speech, and Language Processing*, 21(11): 2379-2392.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2013. Predicting emotional responses to long informal text. *IEEE Trans. Affective Computing*, 4(1):106-115.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1-135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP-02*, pages 79-86.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proc. of SemEval-15*, pages 486-495,

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In *Proc. of SemEval-15*, pages 451-463.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161.

Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani. 2013. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In *Proc. of ESSEM-13*.

Kim Schouten and Flavius Frasincar. 2016. Survey on Aspect-Level Sentiment Analysis. *IEEE Trans. Knowledge and Data Engineering*, 28(3):813-830.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267-307.

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In *Proc. of ACII-11*, pages 121-131.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3): 165-210.

Ruifeng Xu, Lin Gui, Jun Xu, Qin Lu, and Kam-Fai Wong. 2015. Cross lingual opinion holder extraction based on multi-kernel SVMs and transfer learning. *World Wide Web*, 18:299-316.

Liang-Chih Yu, Jin Wang, K. Robert Lai and Xuejie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *Proc. of ACL/IJCNLP-15*, pages 788-793.