CIAL at IJCNLP-2017 Task 2: An Ensemble Valence-Arousal Analysis System for Chinese Words and Phrases

Zheng-Wen Lin
ISA, National Tsing Hua University, Taiwan
victorlin12345@gmail.com

Yung-Chun Chang
Graduate Institute of Data Science, Taipei Medical University, Taiwan
changyc@tmu.edu.tw

Chen-Ann Wang
ISA, National Tsing Hua University, Taiwan
openan7@gmail.com

Yu-Lun Hsieh
IIS, Academia Sinica, Taiwan
morphe@iis.sinica.edu.tw

Wen-Lian Hsu
IIS, Academia Sinica, Taiwan
hsu@iis.sinica.edu.tw

Abstract

Sentiment lexicon is very helpful in dimensional sentiment applications. Because of countless Chinese words, developing a method to predict unseen Chinese words is required. The proposed method can handle both words and phrases by using an ADVWeight List for word prediction, which in turn improves our performance at phrase level. The evaluation results demonstrate that our system is effective in dimensional sentiment analysis for Chinese phrases. The Mean Absolute Error (MAE) and Pearson's Correlation Coefficient (PCC) for Valence are 0.723 and 0.835, respectively, and those for Arousal are 0.914 and 0.756, respectively.

1 Introduction

Due to the vigorous development of social media in recent years, more and more user-generated sentiment data have been shared on the Web. It is a useful means to understand the opinion of the masses, which is a major issue for businesses. However, they exist in the forms of comments in a live webcast, opinion sites, or social media, and often contain considerable amount of noise. Such characteristics pose obstacles to those who intend to collect this type of information efficiently. It is the reason why opinion mining has recently become a topic of interest in both academia and business institutions. Sentiment analysis is a type of opinion mining where affective states are represented categorically or by multi-dimensional continuous values (Yu et al., 2015). The categorical approach aims at classifying the sentiment into polarity classes (such as positive, neutral, and negative,) or Ekman's six basic emotions, *i.e.*, anger, happiness, fear, sadness, disgust, and surprise (Ekman, 1992). This approach is extensively studied because it can provide a desirable outcome, which is an overall evaluation of the sentiment in the material that is being analyzed. For instance, a popular form of media in recent years is live webcasting. This kind of applications usually provide viewers with the ability to comment immediately while the stream is live. Categorical sentiment analysis can immediately classify each response as either positive or negative, thus helping the host to quickly summarize every period of their broadcast.

On the other hand, the dimensional approach represents affective states as continuous numerical values in multiple dimensions, such as valence-arousal space (Markus and Kitayama, 1991), 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. According to the two-dimensional representation, any affective state can be represented as a point in the valence-arousal space by determining the degrees of valence and arousal of given words (Wei et al., 2011; Yu et al., 2015) or texts (Kim et al., 2010). Dimen-

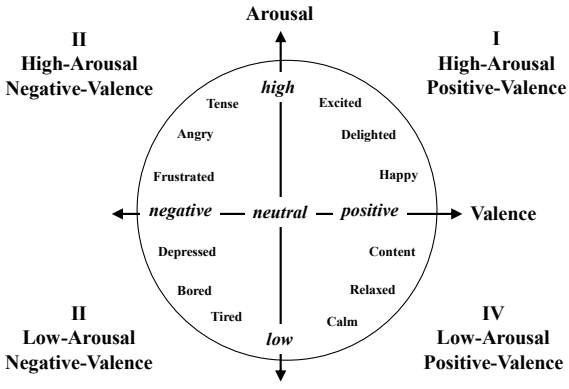

Figure 1: Two-dimensional valence-arousal space.

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 95–99,
Taipei, Taiwan, November 27 – December 1, 2017.

sional sentiment analysis is an increasingly active research field with potential applications including antisocial behavior detection (Munezero et al., 2011) and mood analysis (De Choudhury et al., 2012).

In light of this, the objective of the Dimensional Sentiment Analysis for Chinese Words (DSAW) shared task at the 21th International Conference on Asian Language Processing is to automatically acquire the valence-arousal ratings of Chinese affective words and phrases for compiling Chinese valence-arousal lexicons. The expected output of this task is to predict a real-valued score from 1 to 9 for both valence and arousal dimensions of the given 750 test words and phrases. The score indicates the degree from most negative to most positive for valence, and from most calm to most excited for arousal. The performance is evaluated by calculating mean absolute error and Pearson correlation coefficient between predicted and human-annotated reference scores for two dimensions separately. Participants are required to predict a valence-arousal score for each word, and each phrase.

In order to tackle this problem at the word level, we propose a hybrid approach that integrates valence extension and word embedding-based model with cos similarity to predict valence dimensions. Word embedding-based model with SVM and regression to predict arousal dimensions. At phrase level, we use our ADVWeight List extracted from training sets and our word level method to predict both valence and arousal.

The remainder of this paper is organized as follows. The proposed method is in Section 2. In Section 3, we evaluate performance and compare it with other methods. Finally, some conclusions are listed in Section 4.

2 Method

This study takes 2,802 single words in CVAW 2.0 (Yu et al., 2016) and 2,250 multi-word phrases, both annotated with valence-arousal ratings, as training material. At word level, we use E-HowNet (Chen et al., 2005), a system that is designed for the purpose of automatic semantic composition and decomposition, to extract synonyms of the words from CVAW 2.0, and expand it to 19,611 words with valence-arousal ratings, called WVA. Fig. 2 illustrates the proposed framework. In order to cope with the problem of unknown

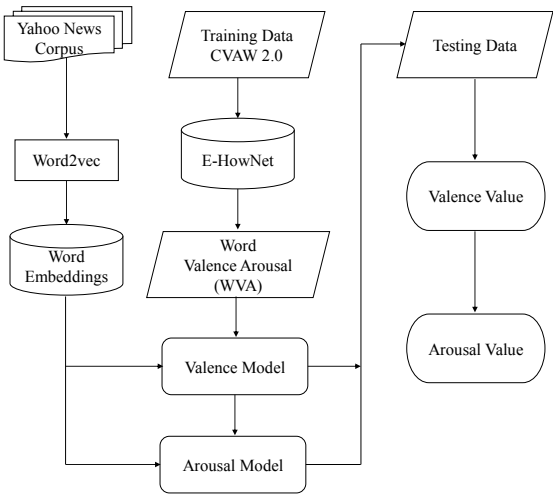

Figure 2: Process of Word Emotional dimension model construction.

words, we separate words in WVA into 4,184 characters with valence-arousal ratings, called CVA. The valence-arousal score of the unknown word can be obtained by averaging the matched CVA. Moreover, previous research suggested that it is possible to improve the performance by aggregating the results of a number of valence-arousal methods (Yu et al., 2015). Thus, we use two sets of methods for the prediction of valence: (1) prediction based on WVA and CVA, and (2) a kNN valence prediction method. The results of these two methods are averaged as the final valence score.

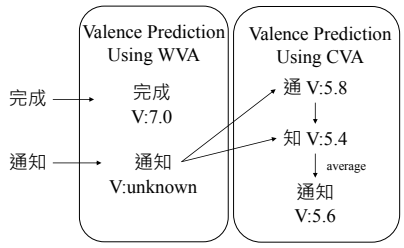

Figure 3: Word Valence prediction method based on WVA and CVA.

First, we describe the prediction of valence values. As shown in Fig. 3, the "完成" of the test data exists in the WVA, so we can directly obtain its valence value of 7.0. However, another word "通知" does not exist in the WVA, so we search in CVA and calculate a valence value of 5.6. Additionally, we propose another prediction method of the valence value, as shown in Fig. 4, based on kNN. We begin by computing the similarity between words using word embeddings(Mikolov

et al., 2013). Then, 10 most similar words are selected and their scores calculated by Eq. 1.

$$\text{Valence}_{\text{KNN}} = \frac{\sum_{i=1}^{x} N_x}{X} \quad (1)$$

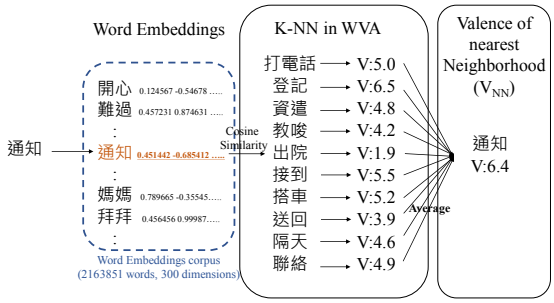

Figure 4: Word valence prediction method based on kNN.

As for the arousal prediction, we propose two methods: (1) linear regression, and (2) support vector regression (SVR) which averages linear regression and SVM predictions as the final arousal score. As shown in Fig. 6, this study considers the linear regression equation in each range according to the valence-arousal value of words in WVA. According to our observation of the data, valence values are generally distributed in the value of 3-7. In order to boost learning of different ranges of data, we distribute them in to two categories. For example, the work "殺害" has a valence value of 1.6. By our design, it will be distributed to categories with valance value of 1 and 2. When the linear regression training is finished, we can predict the corresponding arousal score according to the valence value of the word. As for the SVR-based approach, we first train 300-dimensional word embeddings for all words in WVA using online Chinese news corpus[1]. As shown in Fig. 6, L is the label of the sample, and Dim represents the dimension of the features. We then predict the value of arousal through SVR. Finally, we aggregate the arousal scores predicted by these two methods by taking an average. We observe that the values obtained by linear regression are convergent, while the SVR values are more divergent. So, averaging of the two values can overcome the shortcomings of these methods.

At phrase level, we first experiment with using the proposed word-level model to predict the valence and arousal values. Unfortunately, the results are not satisfactory. We then explore the possibility to incorporate linguistic knowledge into the model. Structurally, phrases can be split into the adverb (ADV) and the adjective (ADJ). An adverb is a word that modifies an adjective in a phrase. For instance, "開心" (happy) with a preceding "非常 (very)" becomes "非常開心 (very happy)," which we consider has an increased degree of happiness. Following this line of thought, we explore ADVs as weighting factors for ADJs. The ADVList and ADVWeight List are extracted from 2,250 multi-word phrases. We employ them to split phrases into ADV and ADJ parts. Subsequently, the valence and arousal values of an ADJ is determined by the word-level prediction model, while those of the ADV is used as an offset. An illustration of our phrase-level prediction process is in Fig. 7.

[1]Collected from Yahoo News between years 2007–2017.

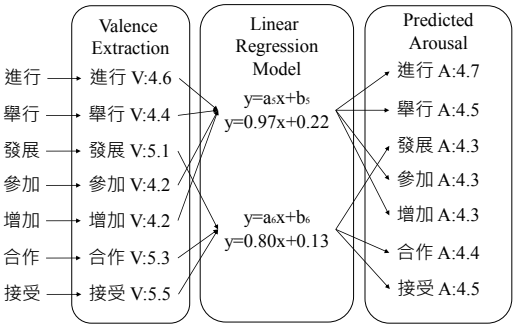

Figure 5: Arousal prediction method based on linear regression.

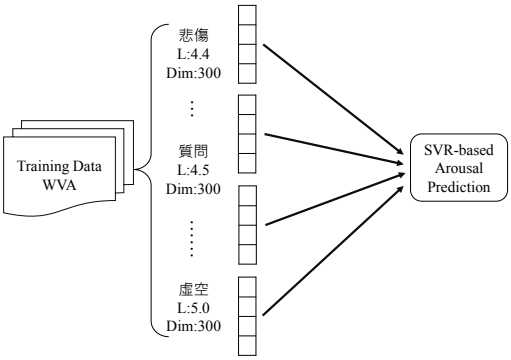

Figure 6: Arousal prediction method based on support vector regression (SVR).

As shown in Fig. 7, in order to obtain the weight of the ADV word "最," we need to use ADVList to split phrases that contain "最" into the format of "[ADV] [ADJ]." Then, our word prediction

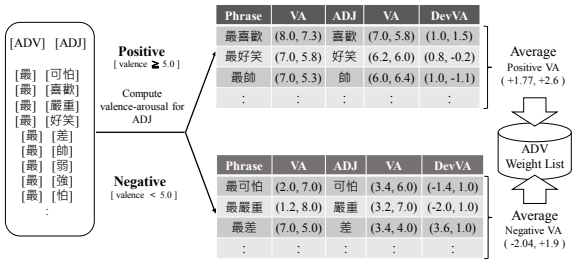

Figure 7: ADV Weight List construction.

model is used to obtain valence (VA) value of the ADJ part. It will be deducted from the VA of the corresponding phrases, and then the remainders are averaged to become the final ADV weight of the word "最". That is, $\mathrm{ADVWeight}(\text{最}) = \mathrm{mean}(\mathrm{VA_{Phrase}} - \mathrm{VA_{ADJ}})$. Most importantly, we hypothesize that ADVs have different effects on phrases with different ADJs, namely, those with valence values ≥ 5.0 and < 5.0. Thus, we have to consider them separately. In the end, there will be four weights for the ADV "最": Positive valence offset, Positive arousal offset, Negative valence offset, and Negative arousal offset.

3 Experiments

We utilize the test data in DSAP_TestW_Input, which contains 750 samples, for performance evaluation. The metrics used are mean absolute error (Mean Absolute Error) and Pearson Correlation Coefficient ($P < 0.01$). In this shared task, valence and arousal values are evaluated separately.

Table 1 shows the results of valence's performance evaluation, $\mathrm{V_{WVA}}$ is the result of WVA alone, and $\mathrm{V_{CVA}}$ is the result of CVA. $\mathrm{V_{WCVA}}$ is the combination of WVA and CVA of the forecast results. $\mathrm{V_{kNN}}$ is a valence prediction method based on kNN. $\mathrm{V_{WVAE}}$ is in WVA through word embeddings to find 10 neighbors, and take the average. Through the comparison of performance, we found that $\mathrm{V_{WCVA}}$ and $\mathrm{V_{WVAE}}$ obtained good results with MAEs being 0.527 and 0.508, respectively, and the PCCs are 0.816 and 0.824. These results suggest that they are highly relevant, so we try to combine the two methods (namely, $\mathrm{V_{mixed}}$.) The final MAE and PCC were 0.496 and 0.845, which is the best-performing method.

Table 2 shows the performance of arousal for different regression methods. $\mathrm{R_{Polyfit}}$ and $\mathrm{R_{Linear}}$ use Polyfit Regression and Linear Regression to predict arousal, while $\mathrm{R_{WVA}}$ is based on linear regression. In addition, $\mathrm{S_{CVAW}}$ and $\mathrm{S_{WVA}}$ use non-corpusated SVR models. $\mathrm{R_{WVA}}$ achieved an outstanding performance of an MAE of 0.939, but was the worst performer in PCC; $\mathrm{S_{WVA}}$ was slightly inferior to $\mathrm{R_{WVA}}$ in MAE, but was superior in PCC with a value of 0.427. Notably, the values predicted by $\mathrm{S_{WVA}}$ are evenly distributed and are more similar to the actual answers, so we try to combine the two methods (RS) to achieve a performance of 0.858 and 0.474 on MAE and PCC, achieving the most outstanding results.

Table 1: Valence method performance.

	$\mathrm{V_{WVA}}$	$\mathrm{V_{CVA}}$	$\mathrm{V_{WCVA}}$	$\mathrm{V_{kNN}}$	$\mathrm{V_{WVAE}}$	$\mathrm{V_{mixed}}$
$\mathrm{MAE_V}$	0.701	0.616	0.527	0.778	0.508	0.496
$\mathrm{PCC_V}$	0.831	0.795	0.816	0.728	0.824	0.845

Table 2: Arousal method performance.

	$\mathrm{R_{Polyfit}}$	$\mathrm{R_{Linear}}$	$\mathrm{R_{WVA}}$	$\mathrm{S_{CVAW}}$	$\mathrm{S_{WVA}}$	RS
$\mathrm{MAE_A}$	1.043	0.953	0.939	1.281	1.003	0.858
$\mathrm{PCC_A}$	0.294	0.296	-0.003	0.367	0.471	0.474

Table 3: Average word-level score and rank of runs 1 and 2 from the participating teams.

Team	$\mathrm{V_{MAE}}$	$\mathrm{V_{PCC}}$	$\mathrm{A_{MAE}}$	$\mathrm{A_{PCC}}$	Rank
AL_I_NLP	0.546	0.8915	0.855	0.6725	1
THU_NGN	0.5595	0.8825	0.9022	0.6545	2
NCTU-NTUT	0.6355	0.844	0.946	0.5545	4
CKIP	0.6335	0.8565	1.041	0.5725	4.5
MainiwayAI	0.7105	0.798	1.0085	0.5305	5.5
CIAL	0.644	0.8515	1.0375	0.4245	6.5
XMUT	0.946	0.701	1.036	0.451	7.5
CASIA	0.725	0.803	1.036	0.451	7.5
Baseline	0.984	0.643	1.031	0.456	8.6
FZU-NLP	1.015	0.645	1.1155	0.4125	10.75
SAM	1.098	0.639	1.027	0.378	10.75
NCYU	1.0785	0.654	1.166	0.415	11.25
NTOU	0.987	0.622	1.1235	0.2565	12.25
NLPSA	1.054	0.5825	1.207	0.351	13.25

Table 3 lists the averaged word-level score and rank of runs 1 and 2 from the participating teams. The Rank column in Table 3 represents the averaged rank of each team. The best-performing team, AL_I_NLP, obtained 0.546 in $\mathrm{V_{MAE}}$, 0.8915 in $\mathrm{V_{PCC}}$, 0.855 in $\mathrm{A_{MAE}}$, and 0.6725 in $\mathrm{A_{PCC}}$. Our method (CIAL) only rank in the middle.

Table 4 lists the averaged phrase-level score and rank of runs 1 and 2 from the participating teams. The Rank column in Table 4 represents the averaged rank of each team. The best-performing team, THU_NGN, obtained 0.347 in

Table 4: Average phrase-level score and rank of runs 1 and 2 from the participating teams.

Team	V_{MAE}	V_{PCC}	A_{MAE}	A_{PCC}	Rank
THU_NGN	0.347	0.9605	0.387	0.91	1
CKIP	0.468	0.928	0.3885	0.906	2.75
NTOU	0.4625	0.9195	0.4305	0.876	3.25
NCTU-NTUT	0.4535	0.9295	0.5025	0.8395	3.5
AL_I_NLP	0.5285	0.9005	0.465	0.8545	4.5
MainiwayAI	0.5945	0.8675	0.539	0.803	6
NLPSA	0.699	0.8235	0.6325	0.7295	7.75
FZU-NLP	0.869	0.697	0.5477	0.785	8
SAM	0.96	0.669	0.722	0.704	10
CIAL	0.9375	0.741	1.255	0.521	11
NCYU	1.105	0.6975	0.768	0.668	11
Baseline	1.051	0.61	0.61	0.61	11
CASIA	1.008	0.598	0.816	0.683	11.5
XMUT	1.723	0.064	1.163	0.084	13.75

V_{MAE}, 0.9695 in V_{PCC}, 0.387 in A_{MAE}, and 0.91 in A_{PCC}. Our method (CIAL) surpasses baseline.

4 Conclusion

The system we developed for DSAW integrates E-HowNet and word embeddings with K-Nearest Neighbors in valence dimension. Support vector regression and linear regression in arousal dimensions. The evaluation results show that the system performance outperforms previous work, but only achieves mediocre performance in this competition. Since the method we used for arousal prediction is still very straightforward, addressing the improvement of its performance should be our target for future research of dimensional sentiment analysis.

Acknowledgments

We are grateful for the constructive comments from three anonymous reviewers. This work was supported by grant MOST106-3114-E-001-002 and MOST105-2221-E-001-008-MY3 from the Ministry of Science and Technology, Taiwan.

References

Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, and Yi-Jun Chen. 2005. Extended-HowNet – a representational framework for concepts. In *Proceedings of OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop*.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media. In *Sixth international AAAI conference on weblogs and social media*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6(3-4):169–200.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, pages 62–70.

Hazel R Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological review* 98(2):224.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Myriam Munezero, Tuomo Kakkonen, and Calkin S Montero. 2011. Towards automatic detection of antisocial behavior from texts. *Sentiment Analysis where AI meets Psychology (SAAIP)* page 20.

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of chinese words from anew. *Affective Computing and Intelligent Interaction* pages 121–131.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K Robert Lai, and Xue-jie Zhang. 2016. Building chinese affective resources in valence-arousal dimensions. In *NAACL/HLT-16*. pages 540–545.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xue-jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *ACL (2)*. pages 788–793.