CKIP at IJCNLP-2017 Task 2: Neural Valence-Arousal Prediction for Phrases

Peng-Hsuan Li
CSIE, National Taiwan University
No. 1, Sec. 4, Roosevelt Rd.
Taipei 10617, Taiwan
jacobvsdanniel@gmail.com

Wei-Yun Ma
IIS, Academia Sinica
No. 128, Sec. 2, Academia Rd.
Taipei 11529, Taiwan
ma@iis.sinica.edu.tw

Hsin-Yang Wang
IIS, Academia Sinica
No. 128, Sec. 2, Academia Rd.
Taipei 11529, Taiwan
whyntut@gmail.com

Abstract

CKIP takes part in solving the Dimensional Sentiment Analysis for Chinese Phrases (DSAP) share task of IJCNLP 2017. This task calls for systems that can predict the valence and the arousal of Chinese phrases, which are real values between 1 and 9. To achieve this, functions mapping Chinese character sequences to real numbers are built by regression techniques. In addition, the CKIP phrase Valence-Arousal (VA) predictor depends on knowledge of modifier words and head words. This includes the types of known modifier words, VA of head words, and distributional semantics of both these words. The predictor took the second place out of 13 teams on phrase VA prediction, with 0.444 MAE and 0.935 PCC on valence, and 0.395 MAE and 0.904 PCC on arousal.

1 Introduction

Sentiment analysis can be a useful tool in understanding public opinions for items of various subjects, such as movies, hotels, and political figures, from unstructured texts. The problem is often defined in two different ways: one that assigns texts to discrete categories, and the other that seeks to get every sample a real value for each dimension (Calvo and Mac Kim, 2013).

For the Dimensional Sentiment Analysis for Chinese Phrases (DSAP) share task of IJCNLP 2017, two dimensions are used to capture the emotions people put in phrases: *valence*, which captures the positive-negative polarity of phrases, and *arousal*, which represents the degree of excitement. The values of both dimensions are limited to a closed interval between 1 and 9, where 1 represents most negative for valence and calmest for arousal.

Type	Count	VA Label
Negation Word	4	No
Modal Word	6	No
Degree Word	42	No
Head Word	2802	Yes
Phrase	2250	Yes

Table 1: Training data statistics for DSAP.

The DSAP shared task calls for systems that automatically predict VA for Chinese phrases to overcome the scarcity of labeled Chinese phrases and words. Lists of words for different types of modifiers are provided. This includes negation words like 不 and 沒有, modal words like 本來 and 應該, and degree words like 有點 and 更加. In addition, some head words with VA annotations are also provided (Yu et al., 2016). Finally, a training data of VA-annotated phrases with their modifier types, e.g. (deg_neg, 稍微不小心), are provided. Table 1 shows the statistics of the training data.

However, besides predicting VA for phrases of which the VA of the head words are known, a seemingly separate task of predicting the VA of unseen words is also required for the competition. Hence, effectively, multiple predictors were built to solve the 4 different problems: phrase valence, phrase arousal, word valence, and word arousal.

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 89–94,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

Hyper-parameter	Trial Range	Final Setting	
		Valence	Arousal
C	$10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}$	10^{1}	10^{1}
ϵ	$10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}$	10^{-2}	10^{-1}
γ	$10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 954^{-1}$	10^{-2}	954^{-1}

Table 2: Grid search hyper-parameters for SVR-RBF.

2 Phrase VA Predictors

Two predictors are constructed and trained similarly for the phrase valence problem and phrase arousal problem. The predictors can be separated into two stages: one that acquires an embedding for a given phrase, and the other that performs the mapping to VA values based on regression analyses.

2.1 Word Segmentation

The first step is to segment a phrase into words. One general way of doing this is to use a popular existing Chinese word segmentation system (Ma and Chen, 2004). However, to best utilize the given knowledge about modifiers and head words with known VA, we developed a simple longest-match segmentation system which uses given data files to make its decision.

For each phrase that are given as a sequence of characters, we first try to match trailing characters to a head word with known VA. Then iteratively leading characters are matched with known modifier words, resulting in a segmented phrase with a sequence of types of its modifiers. For the training data, this matching scheme successfully segments most phrases with correct modifier sequences. The two exceptions are documented below.

First, for the phrase 不是, no head word with known VA can be found. One general solution to this situation is to use word VA predictors to generate the VA of its head word (either 不是 or 是). However, since this phrase is actually not a good sentiment-expressing phrase, we think it is better to simply exclude it from the training data.

The other exception is a set of phrases that ends in 不爽, e.g. 十分不爽. Although both 爽 and 不爽 are words with known VA, 不爽 should be preferred in resolving the segmentation ambiguity according to our longest match principle. However, the resulting modifier type sequence (degree) of (十分,不爽) would be different from the provided (degree,negation) of (十分,不,爽). Recognizing both segmentation can be correct, we choose to split 不 and 爽 as this reduces data sparsity.

2.2 Phrase Features

Having acquired the correct segmentation of phrases, the next question is then how a phrase embedding should be generated for this specific problem. This includes how word embeddings are generated, and how they are combined into phrase embeddings. In addition, some other phrase features that are useful for the problem should be concatenated to these embeddings.

Due to the sparsity of labeled Chinese phrases and words for sentiment analysis, we use unsupervised word embeddings without further tuning. The corpus on which we compute the distributional semantics of Chinese words comes from both the Chinese Gigaword corpus (Graff and Chen, 2003) and the Sinica Corpus (Chen et al., 1996). The former contains over 735 million Chinese characters from the Central News Agency of Taiwan, and the latter contains over 17 million Chinese characters from documents of balanced topics. We use the GloVe algorithm (Pennington et al., 2014) to obtain 300-dimensional word embeddings from a union of these corpora. The resulting 517,015 embeddings cover all words in the training phrases.

To combine word embeddings of phrases to phrase embeddings, we notice the sparsity of available phrase and hence take a simple approach. Observing all given phrases are a compound of one to two modifier words and one head word, we append the word embeddings of the modifier words of a phrase to the word embedding of its head word. With zero paddings to the phrases with only one modifier word, 900-dimensional phrase embeddings are acquired for all phrases.

Finally, two additional features are concatenated to these embeddings. The first is a 2-dimensional VA vector of the head word of each phrase. The second is a 52-dimensional vector indicating which of the 52 modifiers exist in each

phrase. As a result, a 954-dimensional feature vector is composed for every phrase.

2.3 Regression Models

We deploy a series of regression models to gradually approach the problem from the most generalizable models to the most powerful ones, including ridge regression, Support Vector Regression with RBF kernels (SVR-RBF), and multi-layered feed-forward Neural Networks (NN).

The ridge regression is an L2-regularized linear model which is the simplest and fastest because the optimization has an analytical solution. The SVR-RBF adds a non-linear feature transformation before a linear tube regression, leaving many hyper-parameters to be decided but still guaranteeing global optimum for each set of hyper-parameter values. Finally, NN-flavored models are so powerful that every real-valued functions with close-interval domains can be approximated as good as required. However, there is not a guaranteed selection of training schedules of parameters to reach the global optimum.

We acknowledge the sparsity of the labeled data of this task as well as the difficulties in analyzing the non-linear relationships between features and targets of many natural language tasks. Hence, all these models are explored and searched for good hyper-parameters to give an empirical comparison and a suggestion of the best model.

3 Word VA Predictors

As described by Wang and Ma (2016), three predictors are constructed to solve the dimensional sentiment analysis problem for Chinese words.

3.1 E-HowNet-Based Predictor

The first word VA predictor is based on E-HowNet, an expert-built ontology containing the definitions of and relations between about 90 thousand Chinese words. With the knowledge of the sets of synonyms (synsets), the VA of unlabeled words can be predicted by its synonyms of which VA are known.

If multiple labeled synonyms exist for an unlabeled word, the known VA are averaged to give a single prediction. However, if no labeled synonyms of a word can be found, the predictor would fail to predict its VA.

3.2 Word Embedding-Based Predictor

The second predictor relies on distributional semantics of words to determine their similarity. For every unlabeled word, top 10 similar words with known VA are selected, and their VA are averaged as the prediction.

The predictor gains from the fact that most words have pre-trained word embeddings, and hence seldom fails. However, the root cause of failure, the sparsity of labeled words, is not resolved. While the E-HowNet-based predictor gives better results by enforcing synonymity, the embedding-based predictor traded performance for coverage by considering all labeled words in selecting the most similar ones. As a result, the VA of 惡夢 (nightmare) might be used for 美夢 (good dream) because its word embedding is the closest among all labeled words. Averaging the VA of the 10 most similar words other than selecting the most similar one as the prediction somehow alleviates this problem.

3.3 Character-Based Predictor

To enhance the performance of word arousal predictions, a third predictor based on individual characters to propagate labeled arousal is built. The heuristics is that, for many words or even phrases in Chinese, the semantics of their characters contributes strongly to the semantics of the whole. This holds especially when the composing characters are synonyms or near synonyms, e.g. 踴 (leap) and 躍 (jump) for 踴躍 (enthusiastic). Although the contribution is poetic, we could leverage that the words containing similar characters might have similar arousal levels, e.g. 活躍 (active) and 踴躍.

Specifically, the arousal of a character is computed as the average arousal of the labeled words that contains it. Then the arousal of a testing word is predicted as the average arousal of its composing characters which have arousal computed.

4 Experiments

4.1 Phrase Validation Data

The hyper-parameter values of phrase VA predictor models are selected by their performance on the validation set, and the two top-performing predictors are submitted to be evaluated on the testing set. However, as there are only 2250 labeled phrases, 5-fold cross validation is used to gain more reliable evaluations.

Model	Valence MAE	Valence PCC	Arousal MAE	Arousal PCC
Head Word	1.535	0.432	0.794	0.667
Modifier Multiplication	0.522	0.924	0.572	0.836
Ridge	0.967	0.718	0.419	0.898
SVR-RBF	0.408	0.949	0.371	0.919
NN-(750,600,600,450)	**0.334**	**0.966**	**0.361**	**0.922**

Table 3: Cross validation results on training phrases.

Model	Valence MAE	Valence PCC	Arousal MAE	Arousal PCC
Official Linear Baseline	1.051	0.610	0.607	0.730
CKIP-Run1	0.492	0.921	0.382	0.908
CKIP-Run2	0.444	0.935	0.395	0.904
THU_NGN-Run1	0.349	0.960	0.389	0.909
THU_NGN-Run2	**0.345**	**0.961**	**0.385**	**0.911**

Table 4: Testing results of DSAP best submissions.

Specifically, the 2249 segmented phrases, excluding 不是, are randomly shuffled and the first 5 sets of 449 phrases are used as validation samples in turn. All models then share the same 5-fold split of training data.

In addition, we do not group phrases with the same head words, so for example, 也許喜歡, 本來喜歡, and 可能喜歡 might be in different splits. This simulates the fact that unseen phrases might have the same head words as some labeled phrases. However, this could also suffer from overfitting due to data sparsity.

4.2 Baseline Models

Two explainable models are tested to serve as the baseline for the Chinese phrase VA task. The first one, head word model, predicts the VA of a phrase by that of its head word. The second, modifier-multiplication model, multiplies trainable scalar weights of known modifiers to the head word VA (Equation 1 and Equation 2).

$$v_p = 5 + (v_h - 5) \prod_{m \in M} w_m^v \quad (1)$$

$$a_p = 1 + (a_h - 1) \prod_{m \in M} w_m^a \quad (2)$$

p is the testing phrase, h is the head word of p, and M is the set of modifiers of p. v stands for valence, a stands for arousal, and w stands for the trainable weights of each modifier.

The modifier-multiplication model centers head word valence around the median 5, which is presumably the neutrality of opinion polarity. On the other hand, head word arousal are centered around 1, assuming it stands for no excitement. Note that the models degenerate to the head word baseline when all modifier weights default to 1.

Table 3 shows the validation results of the baseline models as well as other models. It can be seen that the multiplication model serves as a strong and explainable baseline.

4.3 Phrase VA Models

Ridge

The ridge regression model has one hyper-parameter: the regularization weight. However, we leave it to be decided by a leave-one-out cross validation of the training split. This gives a non-parametric ridge regression model. As shown in Table 3, it performs worse than the strong baseline on valence but better on arousal.

SVR-RBF

The SVR-RBF model has three hyper-parameters: the error parameter C, the tube parameter ϵ, and the RBF kernel parameter γ. Table 2 shows the trial range of our grid search and the selected best set of values. This non-linear model brings a significant improvement.

NN

Our feed-forward neural networks have a fix L2-regularzation weight of 1. However, we set the possible number of hidden layers to include 1 to 4, and possible dimensions for each layer to include 150, 300, 450, 600, and 750. With a constraint that a layer cannot have a higher dimension

SVR-RBF	Valence MAE	Valence PCC	Arousal MAE	Arousal PCC
h, m, va, mod	0.41	0.95	0.37	**0.92**
m, va, mod	**0.36**	**0.96**	**0.36**	**0.92**
va, mod	0.45	0.93	0.40	0.90
va	1.34	0.44	0.73	0.67
mod	1.31	0.35	0.71	0.66
NN-(300,300,300)	**Valence MAE**	**Valence PCC**	**Arousal MAE**	**Arousal PCC**
h, m, va, mod	**0.34**	**0.97**	**0.36**	**0.92**
m, va, mod	0.37	0.96	0.38	0.91

Table 5: Feature analysis results by cross validation. h stands for the 300d head word embedding, m stands for the 600d modifier embeddings, va stands for the 2d head word VA, and mod stands for the 52d modifier existence vector.

than its previous layer, this yields a total of 125 permutations of network shapes, which will easily explode were a few more dimensions and layers added. Table 3 shows the best configuration, which surpasses other simpler models.

4.4 Phrase VA Test Results

The DSAP shared task releases 750 phrases as the testing set. These phrases have neither VA labels nor modifier information. We use the approaches described in Section 2.1 and Section 2.2 to segment the testing phrases and obtain their 954-dimensional feature vectors.

We submitted the predictions of SVR-RBF and NN, the second and the best performing model in cross validation, as CKIP-Run1 and CKIP-Run2. Our NN model turns out to be one of the best phrase VA predictors, second only to the submissions of team THU_NGN. Table 4 shows these results as well as the official baseline performance of a linear model on word embeddings.

4.5 Phrase VA Feature Analysis

We perform an ablation analysis to shed light on the contributions of each features that lead to highly correlated outputs to the ground truth labels. Table 5 shows the results on cross validation using decreasingly less features. It turns out that just the head word VA plus the information of modifier existence is enough to make highly correlated deductions of phrase VA (above 0.9 PCC), but no single one of them would do. In addition, the contribution of head word embeddings seems to be weak.

4.6 Word VA Test Results

As in the Dimensional Sentiment Analysis of Chinese Words (DSAW) shared task of IALP 2016, An ensemble of the three predictors, the E-HowNet-based, the embedding-based, and the character-based, is used to generate the final submission of testing results. A simple 5:5 ensemble between the E-HowNet-based predictor and the embedding-based predictor (before adding the character-based predictor for arousal) turns out to be one of the best predictors, second only to the submissions of team THU_NGN and team AL_I_NLP. Specifically, 0.602 MAE and 0.858 PCC are achieved for word valence, and 0.949 MAE and 0.576 PCC are achieved for word arousal.

5 Conclusion

We have demonstrated the approaches behind the submissions of team CKIP on the DSAP shared task, and the results on the testing set shows that they are suitable for the task. For the prediction of the valence and the arousal of Chinese phrases, our feature analysis indicates that the non-linear relations between the VA of the head word and the information of modifier appearances are enough to produce highly correlated results to the ground truth. For the prediction of the valence and the arousal of Chinese words, the E-HowNet ontology shows its usefulness again.

The approaches as a whole achieve compelling results for future takes on Chinese sentiment analysis problems, which are expected to be more sophisticated and toward real-world applications.

References

Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in Text: Dimensional and Categorical Models. *Computational Intelligence*, 29(3):527–543.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. *Language*, 167:176.

David Graff and Ke Chen. 2003. Chinese Gigaword LDC2003T09. *Philadelphia: Linguistic Data Consortium*.

Wei-Yun Ma and Keh-Jiann Chen. 2004. Design of CKIP Chinese Word Segmentation System. *International Journal of Asian Language Processing*, 14(3):235–249.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543.

Hsin-Yang Wang and Wei-Yun Ma. 2016. CKIP Valence-Arousal Predictor for IALP 2016 Shared Task. In *International Conference on Asian Language Processing*, pages 164–167.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese Affective Resources in Valence-Arousal Dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics.