NCTU-NTUT at IJCNLP-2017 Task 2: Deep Phrase Embedding using Bi-LSTMs for Valence-Arousal Ratings Prediction of Chinese Phrases

Yen-Hsuan Lee, Han-Yun Yeh and Yih-Ru Wang
Department of Electrical Engineering,
National Chiao Tung University
Hsinchu, Taiwan, ROC
yhl0305.cm05g@g2.nctu.edu.tw
henry034.cm05g@g2.nctu.edu.tw
yrwang@mail.nctu.edu.tw

Yuan-Fu Liao
Department of Electronic Engineering,
National Taipei University of Technology
Taipei, Taiwan, ROC
yfliao@ntut.edu.tw

Abstract

In this paper, a deep phrase embedding approach using bi-directional long short-term memory (Bi-LSTM) neural networks is proposed to predict the valence-arousal ratings of Chinese phrases. It adopts a Chinese word segmentation frontend, a local order-aware word-, a global phrase-embedding representations and a deep regression neural network (DRNN) model. The performance of the proposed method was benchmarked on the IJCNLP 2017 shared task 2. According the official evaluation results, our system achieved mean rank 6.5 among all 24 submissions.

1 Introduction

Recently, sentiment analysis has been approached from many different views. Among them, the two-dimensional valence-arousal space (Yu, et al., 2016) (as in Fig. 1) is promising. In this framework, the valence dimension describes the degree to which an emotion is pleasant or unpleasant, and the arousal dimension describes the degree to which an emotion is associated with high or low energy.

To promote this framework to Chinese language, a series of dimensional sentiment analysis shared tasks (Yu, et al., 2016) had been established since 2016. This paper reports our entry to the second one, i.e., the IJCNLP 2017 Shared Task: Dimensional Sentiment Analysis for Chinese Phrases[1].

Different from previous round, this year's task takes the ratings of phrases into considered. Unlike words, phrases usually have adverbs that could modify or even turn over the sentiment words in different degrees. For example, the word "爽" by itself describes a person in a pleasure mood. And the phrase "好 爽" means the man is feeling so good. But the phrase "不 爽" in fact indicates that person is unhappy. Moreover, the order of words is critical. Comparing the two phrases "完全 不 同意" and "不 完全 同意", the first one means "totally disagree", but the latter one to some extent represents "agree".

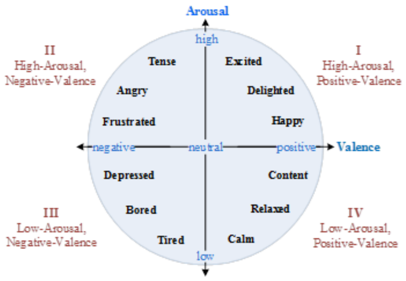

Figure 1: The two-dimensional valence-arousal (VA) space (from [1]).

In order to handle both the modifiers and word order issues, we refresh our previous model (Chou, et al., 2016) and proposed a new deep phrase embedding approach in this paper. The main idea is to build both a local order-aware word- and a recurrent neural network (RNN)-based phrase-embedding representations. To this end, a VA prediction system as shown in Fig. 2 is

[1] http://nlp.innobic.yzu.edu.tw/tasks/dsa_p/

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 124–129,
Taipei, Taiwan, November 27 – December 1, 2017. ©2017 AFNLP

proposed. It adopts a Chinese word segmentation frontend, a local order-aware Word2Vec, a Bi-LSTM Phrase2Vec and a DRNN module.

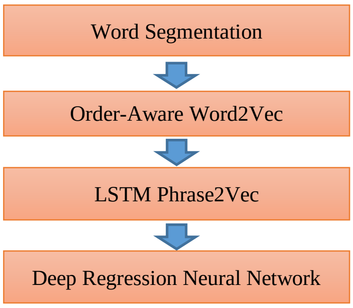

Figure 2: The Block Diagram of the proposed VA ratings prediction of Chinese Phrases approach.

It is worth mentioning that, the purpose of this study is not only to build an effective VA ratings prediction of Chinese phrases system but also to test the performance of our condition random field (CRF)-based Chinese parser (Wang and Liao, 2012). Because a good Chinese word segmentation frontend plays an important role in the success of this task.

2 The Proposed VA Prediction System

The procedures to build the proposed VA ratings system are shown in Fig. 3. First of all, a CRF-based Chinese parser is applied for word segmentation. Then an order-aware Word2Vec and a Phrase2Vec model are trained. Finally, a regression model is adopted to predict the VA ratings of Chinese phrases using the extracted phrase embeddings.

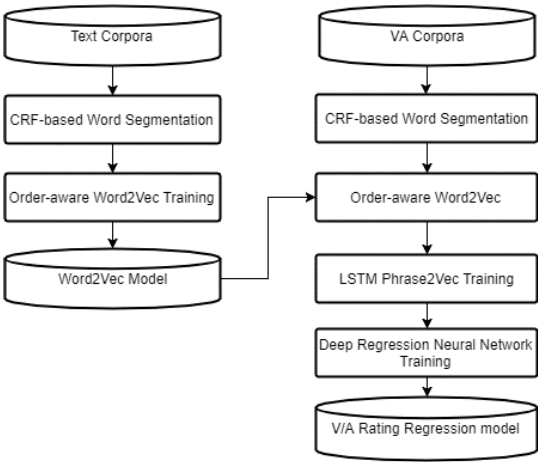
Text Corpora
CRF-based Word Segmentation
Order-aware Word2Vec Training
Word2Vec Model
VA Corpora
CRF-based Word Segmentation
Order-aware Word2Vec
LSTM Phrase2Vec Training
Deep Regression Neural Network Training
V/A Rating Regression model

Figure 3: The procedure flowchart to build the proposed system for VA ratings prediction of Chinese phrases.

In the following subsections, four major system modules will be described in more detail including the (1) CRF-based Chinese parser, (2) order-aware Word2Vec, (3) Bi-LSTM Phrase2Vec and (4) DRNN models.

2.1 Chinese Word Segmentation

Our Chinese word segmentation frontend is a CRF-based parser as shown in Fig. 4. There are three main modules in this system, including (1) text normalization, (2) word segmentation and (3) part of speech (POS) tagging. The whole system was trained using error-corrected version of Sinica Balanced Corpus ver. 4.0[2].

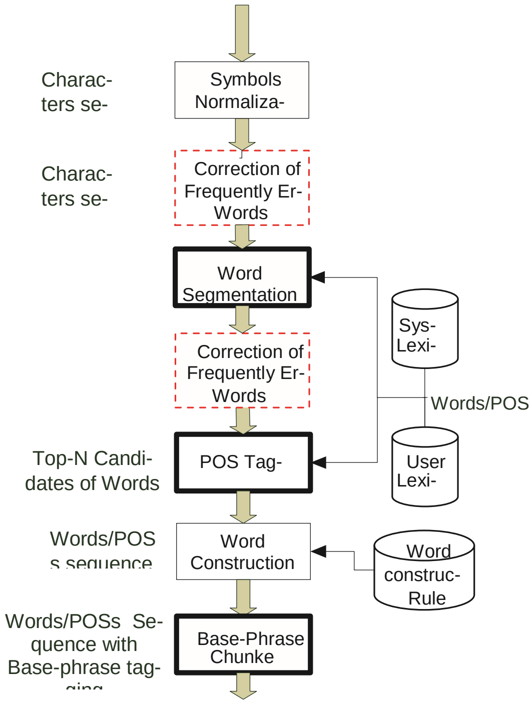

Figure 4: The schematic diagram of the CRF-based Chinese parser.

2.2 Order-Aware Word2Vec

Word to vector (Word2Vec) algorithms (Mikolov, et al., 2013), such as Skip-gram and continuous bag of words (CBOW), are widely used in natural language processing tasks. However, Skip-gram and CBOW only consider the context words but ignore their positions in a phrase. The Consequence is that they cannot deal well with the word order issue.

[2] http://www.aclclp.org.tw/use_asbc.php

To solve this problem, we adopted a modified version of the Continuous Window (CWindow) and Structured Skip-gram approaches proposed by Wang (2015) to preserve the order cues, i.e., the 2-Bag-of-Words (2-BOW) and 2-Skip-gram methods as shown in Fig. 5.

Unlike conventional CBOW and Skip-gram models, 2-BOW and 2-Skip-gram take two contextual bags of words (before and after the target word) into consideration and use two separate projection matrices to preserve the word order information. By this way, order-aware word embedding representations could be generated.

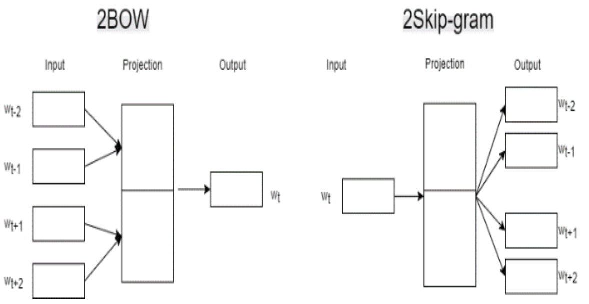

Figure 5: The architectures of the 2-BOW (left) and 2-Skip-gram (right) models.

2.3 Bi-LSTM-based Phrase2Vec

The other difficulty is that the length of phrases is variable. To deal with this problem, the many-to-one LSTM-based (Sepp and Schmidhuber) Phrase2Vec model as shown in Fig. 6 is proposed here.

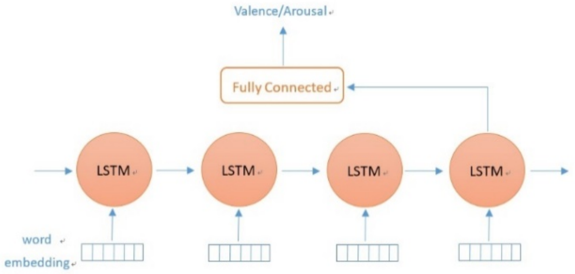

Figure 6: The architecture of the many-to-one LSTM-based Phrase2Vec model.

This is a sequence-to-sequence approach. Each word in a phrase is fed into the LSTM one-by-one at different time step and the last LSTM state vector is treated as the embedding representation of the whole phrase. Using this model, phrases with different numbers of words could all be successfully processed in the same way.

Moreover, a Bi-LSTM-based (Schuster and Paliwal) Phrase2Vec model (as shown in Fig. 7) is also introduced in this paper. The purpose is to explore the word order cues in both the forward and backward directions.

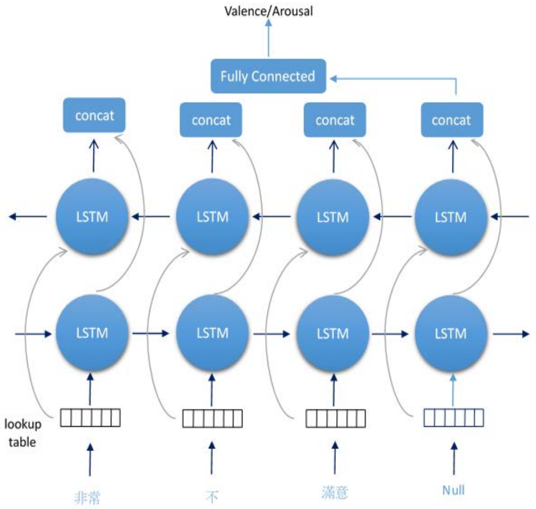

Figure 7: The architecture of the bi-directional LSTM-based Phrase2Vec model.

2.4 Deep Regression Neural Network

Finally, a fully connected feedforward network as shown in Fig. 8 is applied as a non-linear regression model to predict the VA ratings of Chinese phrases.

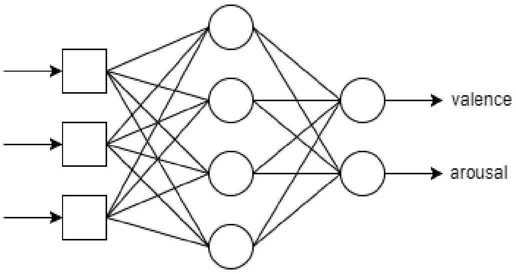

Figure 8: The deep regression neural network for VA ratings prediction of Chinese phrases.

3 Sentiment Analysis Experiments

To benchmark the proposed methods, several preliminary experiments were first conducted to find the optimal model configurations. Then two sets of VA ratings prediction results generated by the best models were submitted for official evaluation.

In the following subsections, the performance of different model combinations will be described in detail.

3.1 Experimental Settings

3.1.1 Text Corpora for Word Embeddings

In this paper, the Skip-gram, CBOW, 2-Skip-gram and 2-BOW models were all trained using the same set of text corpora including (1) LDC Chinese Gigaword Second Edition[3], (2) Sinica Balanced Corpus ver. 4.0, (3) CIRB0303[4], (Chinese Information Retrieval Benchmark, version 3.03), (4) Taiwan Panorama Magazine[5], (5) TCC300[6] and (6) Wikipedia (ZH_TW version).

They were then utilized to project every Chinese word into a high dimensional vector space for further VA analysis.

3.1.2 VA Corpora

The Chinese Valence-Arousal Words 2.0 (CVAW 2.0) and Chinese Valence-Arousal Phrase (CVAP) databases provided by IJCNLP-2017 shared task were divided into a training and a test subsets and used in all the following experiments. Among them, CVAW 2.0 consists of 2,802 affective words and CVAP has 2,250 multi-words phrases. They are all annotated with valence-arousal ratings by hand. In all the following experiments, 10% of CVAW 2.0 and CVAP data are used for validation.

3.1.3 Official Evaluation Metrics

Two evaluation metrics were adopted by the IJCNLP-2017 shared task. The first one is the mean absolute error (MAE, as defined in Eq. (1), lower is better).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - P_i| \quad \ldots\ldots (1)$$

Here A_i denotes the VA ratings ground-truths annotated by human, P_i is automatically predicted VA rating values, n is the number of test samples.

Another one is the Pearson correlation coefficient (PCC, as shown in Eq. (2), higher is better)

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{A_i - \bar{A}}{\sigma_A}\right)\left(\frac{P_i - \bar{P}}{\sigma_P}\right) \quad \ldots\ldots (2)$$

Here $\{\bar{A}, \sigma_A\}$ and $\{\bar{P}, \sigma_P\}$ represent the means and the standard deviations of the human-annotated and machine predicted VA ratings, respectively.

3.2 Preliminary Experimental Results

3.2.1 Different Word Embedding Models

Since the performance of sentiment analysis highly depends on the quality of word embeddings, five different word embedding models including: Skip-gram, CBOW, 2-Skip-gram, 2-CBOW and 2-Skip-gram+2-CBOW were first tested using the same variable input length LSTM (vLSTM) Phrase2Vec backend. In all the experiments, cross validation approach was used to train and test the models. And the window size was all fixed at 15.

Table 1 shows the performances of different Word2Vec models. It was found that 2-Skip-gram+2-BOW and 2-BOW approaches with 300-dimensional embedding vectors achieved the best prediction results for words and phrases, respectively.

Table 1: Performance (MAE) comparison of VA ratings prediction of Chinese words and phrases on different Word2Vec models.

Word	Dimension		
	200	250	300
Skip-gram	0.8275	0.845	0.836
CBOW	0.819	0.8165	0.8205
2-Skip-gram	0.839	0.843	0.8185
2-BOW	0.8085	0.8035	0.8615
2-Skip-gram+2-BOW	0.8035	0.812	**0.783**
Phrase	Dimension		
	200	250	300
Skip-gram	0.541	0.5105	0.491
CBOW	0.4075	0.4055	0.4025
2-Skip-gram	0.4965	0.46	0.458
2-BOW	0.4335	0.4105	**0.3965**
2-Skip-gram+2-BOW	0.397	0.4005	0.407

3.2.2 Different LSTM Structures

Four types of LSTM models were then trained using the same cross validation approaches, including:

- Fixed length LSTM (LSTM): the input sequence was word-padded to the same length.
- Variable length LSTM (vLSTM): The length input sequence is dynamic without word-padding.
- Bi-directional LSTM (Bi-LSTM): Fixed input length but with bi-directional LSTM model.

[3] https://catalog.ldc.upenn.edu/LDC2005T14

[4] http://www.aclclp.org.tw/use_cir.php

[5] https://www.taiwan-panorama.com/en

[6] http://www.aclclp.org.tw/use_mat.php#tcc300edu

- Variable length bi-directional LSTM (vBi-LSTM): Dynamic sequence length plus the bi-directional LSTM model.

Table 2 reports the performance of different LSTM models (all with the 2-Skip-gram+2BOW with 300-dimension word embeddings frontend). The results indicate that, in general, the variable input length LSTM models work better than the fixed ones.

Table 2: Performance (MAE) comparison of VA ratings prediction of Chinese words and phrases on different LSTM models.

Word	Valence	Arousal	Average
LSTM	0.64	0.954	0.797
vLSTM	0.625	0.982	0.8035
Bi-LSTM	**0.622**	0.961	0.7915
vBi-LSTM	0.638	**0.928**	**0.783**
Phrase	Valence	Arousal	Average
LSTM	0.377	0.415	0.396
vLSTM	**0.364**	0.418	**0.391**
Bi-LSTM	0.387	**0.402**	0.3945
vBi-LSTM	0.388	0.426	0.407

3.3 Official Evaluation Results

Based on the results of the preliminary experiments, two runs (NCTU+NTUT run1 and run2) were submitted to ICJNLP2017 shared task for official benchmark. Both run1 and run2 adopted the 2-Skip-gram+2-BOW model with 300-dimension word embeddings plus the variable input length LSTM models. The only difference is that run1 used the vLSTM and run2 adopted the vBi-LSTM model, respectively.

Table 3 and 4 shows the official evaluation results of our two submissions. The performance of the run1 and run2 are all promising (with only a marginal difference between these two models).

Table 3: Official MAE evaluation results of the NCTU+NTUT's submissions (Run1 and Run2).

Word	Valence	Arousal	Average
Run1	**0.632**	0.952	0.792
Run2	0.639	**0.94**	**0.7895**
Phrase	Valence	Arousal	Average
Run1	0.454	**0.488**	**0.471**
Run2	**0.453**	0.517	0.485

Table 4: Official PCC evaluation results of the NCTU+NTUT's submissions (Run1 and Run2).

Word	Valence	Arousal	Average
Run1	**0.846**	0.543	0.6945
Run2	0.842	**0.566**	**0.704**
Phrase	Valence	Arousal	Average
Run1	0.928	**0.847**	**0.8875**
Run2	**0.931**	0.832	0.8815

4 Conclusions

In this paper, we had proposed and evaluated various order-aware embedding representations in both word- and phrase-levels. It is found that the order-aware Word2Vec and LSTM-based Phrase2Vec all could improve the performance of VA ratings prediction of Chinese phrases. In brief, our system achieved mean rank 6.5 among in total 24 submissions in the official ICJNLP2017 shared task evaluation. Finally, the latest version of our Chinese parser is available on-line at http://parser.speech.cm.nctu.edu.tw/.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan with contract 105-2221-E-027-119, 105-2221-E-009-142-MY2 and 106-2221-E-027-128.

References

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

Liang-Chih Yu, Lung-Hao Lee and Kam-Fai Wong. 2016b. Overview of the IALP 2016 shared task on dimensional sentiment analysis for Chinese words. In *IALP*, pages 156-160.

Wei-Chieh Chou, Chin-Kui Lin, Yih-Ru Wang, Yuan-Fu Liao. 2016. Evaluation of weighted graph and neural network models on predicting the valence-arousal ratings of Chinese words. In *IALP* pages 168-171.

Yih-Ru Wang and Yuan-Fu Liao. 2012. A Conditional Random Field-based Traditional Chinese Base-Phrase Parser for SIGHAN Bake-off 2012 Evaluation. In *CLP 2012*. pages 231.

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.

Ling Wang, Chris Dyer, Alan Black and Isabel Trancoso. 2015. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In *HLT-NAACL*.

Hochreiter Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural computation 9.8,* pages 1735-1780.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. In *IEEE Transactions on Signal Processing* 45.11. pages 2673-2681