IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 November 2022, accepted 28 November 2022, date of publication 1 December 2022,
date of current version 7 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3226243

RESEARCH ARTICLE

Predicting Chinese Phrase-Level Sentiment Intensity in Valence-Arousal Dimensions With Linguistic Dependency Features

YU-CHIH DENG[1], CHENG-YU TSAI[1], YIH-RU WANG[2], SIN-HORNG CHEN[1], AND LUNG-HAO LEE[3], (Member, IEEE)

[1]Department of Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[2]Department of Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan *(Retired)*
[3]Department of Electrical Engineering, National Central University, Taoyuan 320317, Taiwan

Corresponding author: Lung-Hao Lee (lhlee@ee.ncu.edu.tw)

This work was supported in part by the ASUS Inc. with Taiwan Computing Cloud (TWCC) service, and in part by the National Science and Technology Council, Taiwan, under Grant MOST 111-2628-E-008-005-MY3 and Grant MOST 108-2218-E-008-017-MY3.

ABSTRACT Phrase-level sentiment intensity prediction is difficult due to the inclusion of linguistic modifiers (e.g., negators, degree adverbs, and modals) potentially resulting in an intensity shift or polarity reversal for the modified words. This study develops a graph-based Chinese parser based on the deep biaffine attention model to obtain dependency structures and relations. These obtained dependency features are then used in our proposed Weighted-sum Tree GRU network to predict phrase-level sentiment intensity in the valence-arousal dimensions. Dependency parsing results using the Sinica Treebank indicate that our graph-based model outperforms transition-based methods such as MLP and stack-LSTM with identical findings for English dependency parsing. Experimental results on the Chinese EmoBank indicate that our Weighted-sum Tree GRU network model outperforms other transformer-based neural networks such as BERT, ALBERT, XLNET and ELECTRA, reflecting the effectiveness of linguistic dependencies in phrase-level sentiment intensity predication tasks. In addition, our proposed model requires fewer parameters and less inference time for quantitative analysis, making the proposed model is relatively lightweight and efficient.

INDEX TERMS Dependency parsing, dimensional sentiment analysis, affective computing, deep learning.

I. INTRODUCTION

Sentiment analysis involves the use of linguistic processing to differentiate the positive and negative emotional content of utterances, as well as their emotional strength values [1], [2], [3]. Continuous real-valued sentiment scores, called 'intensity', provide more fine-grained emotional information. SemEval-2016 Task 7 focused on determining the sentiment intensity of English and Arabic utterances [4]. Various participating teams achieved promising results using different methods includes random forest [5], pointwise mutual information [6], Gaussian regression [7] and linear regression with manual rules [8]. A shared task on dimensional sentiment analysis for Chinese phrases was also organized at IJCNLP-2017 [9]. Affective states were represented in the valence-arousal space [10]. 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. Deep learning-based neural computing approaches such as ensemble Long Short-Term Memory models [11], boosted neural networks [12] and feed-forward neural networks [13] were used to predict the sentiment intensity of Chinese multi-word phrases.

Linguistic modifiers such as negators (e.g., not, never), degree adverbs (e.g., very, totally, slightly) and modals (e.g., would, could) are commonly used in opinion expressions, and play an important role in recognizing sentiment intensity [14]. For example, in Chinese "完全不同意" (totally not agree) and "不完全同意" (not totally agree) convey different meanings. The former is composed of a degree adverb "完全" (totally), a negator "不" (not) and a verb "同意" (degree), meaning that the speaker totally disagrees with the

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

subject, while the latter features the same modifiers in a different order, meaning that the speaker does not completely disagree, but rather partially agrees. Phrase-level sentiment intensity prediction is difficult because linguistic modifiers may lead to an intensity shift or polarity reversal for the words they modify [15].

However, Chinese syntactic parsing to obtain linguistic dependency information is rarely addressed, motivating us to develop a Chinese dependency parser to extract dependency structures and relation features between words for phrase-level sentiment analysis. In addition, Chinese phrases may have the same dependency parsing results such as the previous examples "完全不同意" (totally not agree) and "不完全同意" (not totally agree), but expressing different meanings with almost opposite affective states. Hence, we propose a Weighted-sum Tree Gated Recurrent Unit (Tree GRU) network to tackle the same ordering problem, originating from different sentences with the same dependency structure, for phrase-level sentiment intensity prediction in valence-arousal dimensions.

The main contributions are summarized as follows.

(1) Developing a Chinese Dependency Parser for Syntactic Structure Analysis

We develop a graph-based Chinese dependency parser based on the deep biaffine attention model [16] to obtain linguistic structures and relations between words. The Sinica Treebank [17] was used to evaluate dependency parsing results, indicating that our graph-based model outperforms transition-based methods such as MLP [18] and stack-LSTM [19] with identical findings for English dependency parsing.

(2) Exploring Linguistic Dependency Features for Chinese Phrase-level Sentiment Intensity Prediction

We propose a Weighted-sum Tree GRU network to leverage linguistic dependency features for phrase-level sentiment intensity prediction in the valence-arousal dimensions. Chinese Valence-Arousal Phrases (CVAP) from the Chinese EmoBank corpus [20] were used to evaluate performance. In experiments, our Weighted-sum Tree GRU neural network with linguistic dependency information outperformed other transformer-based neural networks (i.e., BERT [21], RoBERTa [22] and MacBERT [23], ALBERT [24], XLNet [25], and ELECTRA [26]), in the two-dimensional valence-arousal space, confirming the effectiveness of exploited linguistic dependency features. In addition, our proposed model contains fewer parameters and requires less inference time for quantitative analysis, so our proposed model is relatively lightweight and efficient.

The rest of this paper is organized as follows. Section 2 reviews related studies for phrase-level sentiment intensity predication. Section 3 describes our proposed network architecture for valence-arousal rating prediction. Section 4 describes experiments and discusses experimental results for model performance evaluation. Conclusions are drawn in Section 5.

II. RELATED WORK

This section describes existing methods for sentiment intensity prediction for multi-word phrases, including heuristic-based [27], [28], [29], [30], [31], [32], [33], [34] and learning-based [5], [6], [7], [8], [11], [12], [13], [15], [35], [36] methods.

A. HEURISTIC-BASED METHODS

Heuristic-based methods use human-estimated or experience-determined weights to capture the intensity of the modifier's influence on the affective strength for the modified word. Heuristic methods can be further categorized as switch [27], [28], [29], [30] and shift models [27], [28], [29], [30], [31], [32], [33], [34].

Both switch and shift models are used to constrain the negation [27], [28], [29], [30]. A contextual shift approach was proposed to predict positive and negative sentiment for each term [27], [28], incorporating an optional SVM algorithm to learn and classify the sentiment shifts composed of bi-gram and uni-gram features to obtain better classification performance. A rule-based model was used to detect the intensity of emotions in informal English [29], which was improved using an unsupervised version of SentiStrength 2 [30].

Linguistic features are identified based on semantic rules and use a linear offset model to classify sentiment [31]. The Semantic Orientation CALculator (SO-CAL) is applied to the polarity classification task [32], assigning a positive or negative label to a text to capture textual opinions related to the main topic. A linguistic modifiers-based model was proposed to improve emotion classification by designing negation, intensifiers and modalities that may change the emotional meaning of the text [33]. The Valence Aware Dictionary for sEntiment Reasoning (VADER) uses a rule-based model that constructs gold-standard lists through lexicon features [34].

B. LEARNING-BASED METHODS

The learning-based methods use machine learning [5], [6], [7], [8], [35], [36] and deep learning [11], [12], [13], [15] techniques for sentiment intensity prediction.

Machine learning approaches focus on the use of data and algorithms to train models to predict sentiment scores. Random forest [5] was used as a pairwise strategy to predict the sentiment intensity scores. Point-wise mutual information [6] was used to check for similarity between words and prototypical sets, where words with high similarity were incorporated into the emotional lexicon. Adaptive boosting [36] was used to combine multiple weak classifiers into a single strong classifier. The Gaussian regression model [7], [35] was used to compute sentiment intensity scores by incorporating multiple features including direct search results, Word2Vec search, rule-based search, and 5-level Stanford sentiment classifier output [37]. A linear regression model [8] was used to analyze the data noise that affected sentiment intensity prediction performance.

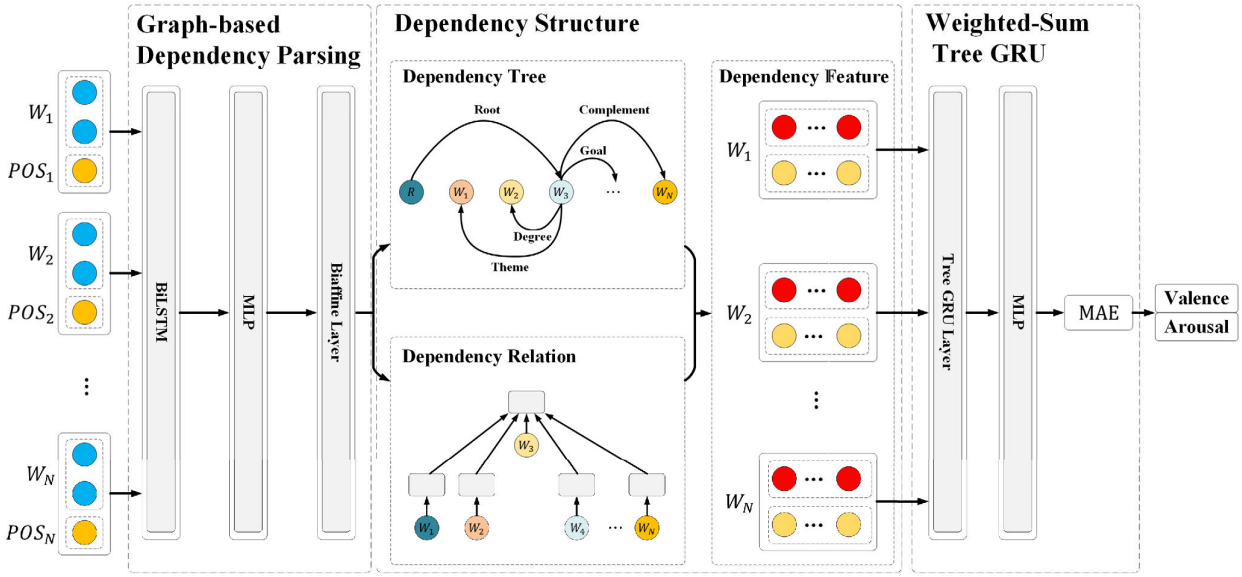


FIGURE 1. Proposed network architecture for phrase-level sentiment intensity prediction.

Deep learning techniques can be used in a variety of ways, including modifying the architecture of neural networks or integrating multiple neural network models. The Part-Of-Speech (POS) embedding and word cluster was fed into the dense Long Short-Term Memory (LSTM) network architecture [11], to undergo 100 training iterations using different hyper-parameters and training data to improve generalization and reduce data noise. A boosted neural network model [12] was used to improve the accuracy of misinterpreted data. A multi-layered feed-forward neural network [13] was proposed to include the types of known modifier words, valence-arousal value of headwords, and the distributional semantics of both kinds of words for valence-arousal intensity prediction. A pipelined neural network model composed of two neural networks (NN) models was proposed to predict phrase-level sentiment intensity [15], in which the first NN model was used to combine the re-weighting mechanism in the hidden layer, and the second NN model considered not only individual but also group weights.

In summary, we follow the research development of neural networks-based deep learning methods since neural computing techniques usually achieve promising results. In this paper, we propose a Weighted-sum Tree GRU network to fully use of exploited dependency features, obtained by our developed Chinese dependency parser, for phrase-level sentiment intensity prediction in the valence-arousal dimensions.

III. CHINESE PHRASE-LEVEL SENTIMENT INTENSITY PREDICTION

Figure 1 shows our proposed network architecture for Chinese phrase-level sentiment intensity prediction, comprised of two main parts: 1) graph-based dependency parsing; and 2) a Weighted-sum Tree GRU network.

A. GRAPH-BASED DEPENDENCY PARSING

We use the graph-based deep biaffine attention model [16] for Chinese dependency parsing. At the embedding layer, the concatenation of a pretrained skip-gram word embedding and a trainable Part-of-Speech (POS) embedding is used as the representation for each word. The recurrent output vector from the following Bidirectional Long Short-Term Memory (BiLSTM) layers then serves as the contextualized word representation for dependency parsing. We then reduce the dimensionality of the recurrent output vector using the Multi-Layer Perceptron (MLP) layers to strip away irrelevant information. Finally, the scores of all the directed arcs between every pair of words are calculated using the biaffine transformation. The cross-entropy loss function is used to calculate the loss at training time, while at testing time the optimal parsing tree is searched using the Maximum Spanning Tree algorithm. Our network architecture uses two deep biaffine attention models with common embeddings and BiLSTM layers are used to obtain dependency arc and type. The objective is to predict the probabilities of all modified words in a sentence. After training the deep biaffine attention model, we can obtain the probability matrix of all arcs and the corresponding dependency matrix. The obtained dependency tree structure and relations will be used respectively as the input order and features in the following Weighted-sum Tree GRU network.

B. WEIGHTED-SUM TREE GRU NETWORK

Considering the use of dependency relations of words and the syntactic information of the dependency tree in our model, we adopt the tree-structured Recurrent Neural Networks (RNNs) [38]. The benefit of the tree-structured RNN over the standard RNN is its capability to compute the hidden states of multiple children from their hidden states. The order of input features to the tree-structured RNN follows the tree

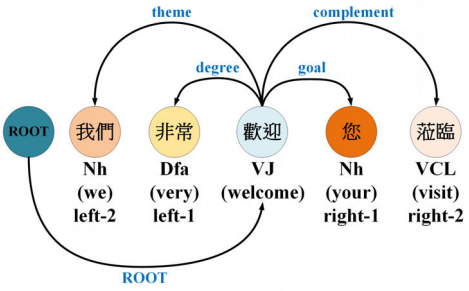


FIGURE 2. Left-right order in our proposed model.

structure of the dependency parsing results (dependency tree and dependency relation). Then, the hidden states of the direct syntactic children of the ROOT nodes are passed to the following feed-forward network to predict VA scores from 1 (highly negative or clam) to 9 (highly positive or excitable).

Different sentences with various degrees of sentiment intensity may share the same dependency tree structure, producing identical features and thus resulting in incorrect VA prediction. To solve this problem, we propose a Weighted-sum Tree GRU network, where the state of a component node state (h) is produced based on the multiple hidden states (h_k) of its children. To model the influence of the different left-right order of the dependents, different weight matrices ($U_k^{(r)}$ for the reset gate: r_k; $U_k^{(h)}$ for the candidate hidden state: $\hat{h}$; $U_k^{(z)}$ for the update gate: z) will be learned for the input hidden states (h_k) of the different left-right order (k). Different sets of weight matrices are used when inputting hidden states into the node state of the headword. The discussion of a headword accounts for the left-right order of the dependency words. Figure 2 shows the left-right order of the sentence "我們非常歡迎您蒞臨"(We very welcome your visit). We set a window size as 2 to minimize the number of weight matrices. For instance, "我們" (we) is the left-2 dependent (k= -2) and "蒞臨" (visit) is the right-2 dependent (k = 2) of their headword "歡迎" (welcome). For the dependent words with a left-right order greater than N, the leftmost Left-N weight matrices ($U_{-N}^{(r)}$, $U_{-N}^{(h)}$, $U_{-N}^{(z)}$) or the rightmost Right-N weight matrices ($U_N^{(r)}$, $U_N^{(h)}$, $U_N^{(z)}$) will be used.

Figure 3 shows the pseudocode of our proposed network architecture. The dependency features and the window index are fed into the node_forward function as the input to calculate the hidden state h, reset gate r and update gate z. The transition equations are described in detail as follows. The component node state h is a linear interpolation between the previous hidden state $\tilde{h}$; and the candidate hidden state $\hat{h}$, as shown in Eq. (1), where the update gate z decides how much the unit updates its previous hidden state, and it is computed by Eq. (2). The previous hidden state is the summation of all the input hidden states shown in Eq. (3). The candidate hidden state is then computed by Eq. (4), where the reset gate r is computed as Eq. (5). When calculating the candidate hidden state, the previous hidden state (h_k) state is ignored if the corresponding reset gate (r_k) is close to 0.

```
def node_forward(self, inputs, child_h, windows_idx):
    for i in range(len(window_idx)):
        reset_h = (update_h[window_idx[i] + lr_window](child_h[i])
    reset = sigmoid(reset_h + reset_x)      # Equation (2), (5)
```

$$z = \sigma(W^{(z)}\mathbf{x} + \sum_k U_k^{(z)} \cdot h_k) \quad \forall k \le -N, U_k^{(z)} = U_{-N}^{(z)}; \forall k \ge N, U_k^{(z)} = U_N^{(z)} \tag{2}$$

$$r_k = \sigma(W^{(r)}\mathbf{x} + U_k^{(r)} h_k) \quad \forall k \le -N, U_k^{(r)} = U_{-N}^{(r)}; \forall k \ge N, U_k^{(r)} = U_N^{(r)} \tag{5}$$

```
    reset = mul(reset, child_h)
    for i in range(len(window_idx)):
        u = u + uh[window_idx[i] + lr_window](reset[i])
    u = tanh(u)      # Equation (4)
```

$$\hat{h} = \tanh(W^{(h)}\mathbf{x} + \sum_k U_k^{(h)} \cdot r_k \cdot h_k) \quad \forall k \le -N, U_k^{(h)} = U_{-N}^{(h)}; \forall k \ge N, U_k^{(h)} = U_N^{(h)} \tag{4}$$

```
    lr_window = 2  # window size
    update = mlp_layer(inputs)
    for i in range(len(window_idx)):
        update = update + update[window_idx[i] + lr_window](child_h[i])
    child_h_avg = sum(child_h) / 2      # Equation (3)
```

$$\tilde{h} = \sum_k h_k \tag{3}$$

```
    h = mul(child_h_avg, update) + mul(u, 1-update)      # Equation (1)
```

$$h = z \cdot \tilde{h} + (1 - z) \cdot \hat{h} \tag{1}$$

```
def forward(self, tree, inputs):
    Training initialization()
    # Weighted-sum Tree GRU network
    tree.state = node_forward(inputs, child_h, window_idx)
    va_pred = mlp_va(sigmoid(mlp_layer(tree.state)))
```

FIGURE 3. Pseudocode for weighted-sum tree GRU network.

Finally, the predicted valence and arousal values are outputted after passing through the MLP.

Compared with the child-sum Tree GRU network [39], our proposed network imports the past h_k information of various lower-level nodes when computing the node state h of the headword. However, the previous hidden-layer states of all modifiers for this headword are treated using the same parameters. Therefore, the child-sum Tree GRU model [39] cannot simulate the ordering of the target word in a sentence, mainly because different sentences sometimes generate identical dependency tree structures. Figure 4 uses the two phrases "完全不同意" (totally not agree) and "不完全同意" (not totally agree) as examples, expressing different meanings with almost opposite affective states, but having the same dependency parsing result. Hence, these two phrases have the same valence-arousal rating prediction in the child-sum Tree GRU framework. Our proposed Weighted-sum Tree GRU network for the sentiment intensity task can handle the same ordering problem. We use a set of weight matrices to reflect

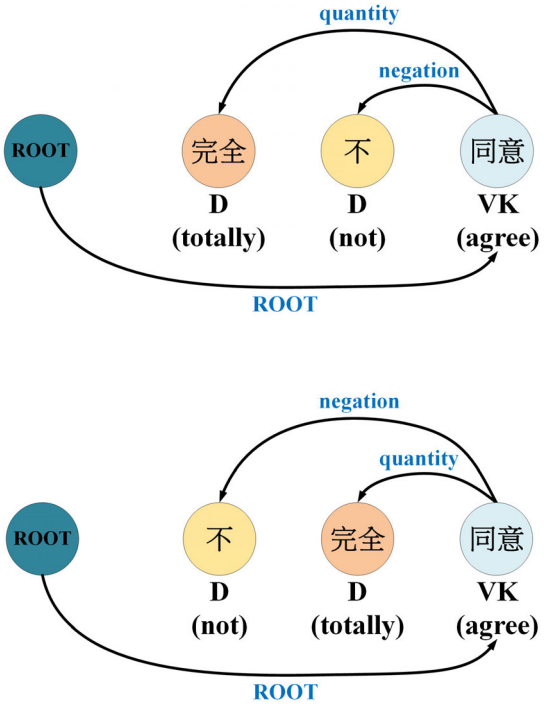


FIGURE 4. Different sentences with the same dependency tree.

the modifier hidden state information. These weight matrices are learned using the Tree GRU neural network to extract the related reset gate, candidate hidden state and update gate features of the input phrases, allowing the proposed method to handle different sentences with the same dependency structure.

IV. EXPERIMENTS FOR PERFORMANCE EVALUATION

A. DATASETS

Sinica Treebank [17] was divided into two mutually exclusive datasets to evaluate dependency parsing performance. The training set includes 56,957 sentences with a total of 337,174 words. The test set includes 690 sentences with 5,160 words.

The Chinese Valence-Arousal Phrases (CVAP) set from the Chinese EmoBank [20] was used to evaluate sentiment intensity prediction performance. A total of 52 modifiers (including 4 negators, 42 degree adverbs, and 6 modals) were combined with the affective words in the Chinese Valence-Arousal Words (CVAW) set [40] to form multi-word phrases. VA ratings were annotated through crowdsourcing with each phrase randomly assigned to 10 annotators. Both the valence and arousal dimensions use a nine-degree scale. A value of 1 on the valence and arousal dimensions respectively denotes extremely high-negative and low-arousal sentiment, while a 9 denotes extremely high-positive and high-arousal sentiment, and 5 denotes a neutral and medium-arousal statement. Outlier ratings were identified and excluded from the calculation of the average VA ratings for each phrase. Finally, a total of 2,998 Chinese phrases were constructed in the CVAP. We randomly distributed in groups of 5 for cross-validation evaluation.

B. SETTINGS

Embedding training was performed using the following corpora: Chinese Gigaword (Ver. 2.0),[1] Sinica Corpus (Version 4.0),[2] Chinese Information Retrieval Benchmark (Version 3.03),[3] Taiwan Panorama Magazine,[4] Microphone Speech Database (TCC300),[5] and Chinese Wikipedia.[6] We used an automatic system [41] to obtain the segmented words and their corresponding part-of-speech tags to train Word2Vec vectors [42].

All experiments were implemented using PyTorch. The hyper-parameters of our proposed Weighted-sum Tree GRU network were set up as follows: batch size 256; word vector dimension 250; POS vector dimension 50; parameter dimension of dependency features was 100; memory size of Weighed-sum Tree GRU was 256; hidden state of MLP was 512; Adagrad was used as the optimizer; and the number of epochs was restricted to 50. We use BERT,[7] RoBERTa,[8] MacBERT,[9] ALBERT,[10] XLNet,[11] and ELECTRA[12] to compare the performance of sentiment intensity prediction with the following hyper-parameters: batch size 64; average pooling style; the pre-trained models with 12-layer, 768-hidden and 12-heads; the optimizer is AdamW; and the number of epochs was 20.

C. METRICS

Two metrics were used to evaluate paring results: 1) Unlabeled Attachment Score (UAS), the proportion of tokens that are assigned the correct head, and 2) Labeled Attachment Score (LAS), the proportion of tokens that are assigned both the correct head and the correct dependency relation label.

The sentiment intensity predication performance is evaluated by examining the difference between machine-predicted ratings and human-annotated ratings using two metrics to independently evaluate the valence and arousal predictions: Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC), defined as Eq. (1) and (2).

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|a_i - p_i| \tag{1}$$

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_A}\right) \tag{2}$$

where $a_i \in A$ and $p_i \in P$ respectively denote the i-th actual value and predicted value, n is the number of test samples, μ_A

[1] https://catalog.ldc.upenn.edu/LDC2005T14
[2] http://www.aclclp.org.tw/use_asbc.php
[3] http://www.aclclp.org.tw/use_cir.php
[4] https://www.taiwan-panorama.com/
[5] http://www.aclclp.org.tw/use_mat.php#tcc300edu
[6] https://zh.wikipedia.org/wiki
[7] Multilingual BERT: https://github.com/google-research/bert
[8] RoBERTa: https://huggingface.co/hfl/chinese-roberta-wwm-ext
[9] MacBERT: https://huggingface.co/hfl/chinese-macbert-base
[10] Chinese ALBERT: https://github.com/google-research/albert
[11] XLNet: https://huggingface.co/hfl/chinese-xlnet-base
[12] ELECTRA: https://huggingface.co/hfl/chinese-electra-base-discriminator

TABLE 1. Results of dependency parsing.

Model	UAS (%)	LAS (%)
MLP	88.3	83.5
Stack-LSTM	89.8	84.5
Deep Biaffine Attention	**92.9**	**88.5**

and σ_A respectively represent the mean value and the standard deviation of A, while μ_P and σ_p respectively represent the mean value and the standard deviation of p. the mae measures the error rate and the pcc measures the linear correlation between the actual values and the predicted values. A lower MAE and a higher pcc indicate more accurate prediction performance.

D. DEPENDENCY PARSING RESULTS

In the first set of experiments, two transition-based methods were used to compare performance as follows:

- MLP (Multi-layer Perceptron) [18]

An MLP-based fast parser is proposed to obtain dependency parsing results. The input features, including words, POS tags and arc labels are merged and then all feature parameters are summed by linear transformation. Finally, the softmax function is used to make the classification decision.

- Stack-LSTM (Stack Long Short-Term Memory) [19]

A transition-based RNN is proposed to follow the Stanford parser's features from the partial parsing trees, combining the partial dependency tree into the highest two layers of the stack. In addition to words, the dependency tree also contains actions and labels. Therefore, if the properties are different, the labels and words are trained by composition, then the stack is updated by a linear transformation.

Table 1 shows the dependency parsing results. The graph-based method (our adopted Deep Biaffine Attention model) outperforms the MLP [18] and Stack-LSTM [19], with identical findings for English dependency parsing [43]. In our experience, when conducting transition-based methods, words in a sentence are put into the stack from left-to-right. However, sentences can have complicated and non-linear syntactic structures. The graph-based method calculates the weights of all possible edges from word to word and searches for the optimal solution using graph theories, which is more suitable for Chinese syntactic structure.

E. SENTIMENT INTENSITY PREDICTION RESULT

In the second set of experiments, the following transformer-based models were compared to demonstrate their performance for phrase-level sentiment intensity prediction.

- BERT (Bidirectional encoder Representations for transformers) [21]

BERT uses an encoder architecture with an attention mechanism to construct a transformer-based neural network architecture, providing state-of-the-art results in a wide variety of natural language processing tasks. BERT was pre-trained on two tasks: 1) Masked Language Models (MLM): a fixed ratio of tokens was masked to train BERT and the model then predicts the original value of the masked words based on the context; 2) Next Sentence Prediction (NSP): BERT was trained to predict whether the following sentence was probable or not based on the previous sentence. Through pre-training, BERT learns contextual embeddings for representations from large-scale data sets. After pre-training, BERT can be fine-tuned on smaller data sets to optimize its performance on specific tasks.

- RoBERTa (a Robust optimized BERT pre-training approach) [22]

RoBERTa is a replication study of BERT pre-training that carefully measures the impact of key parameters and training data size. The model modifications include removing next sentence predictions, dynamically changing the masking pattern applied to the training data, and training in large batches.

- MacBERT (MLM as correction BERT) [23]

MacBERT revisits the Chinese pre-trained language model series and improves upon RoBERTa, particularly the masking strategy that adopts MLM as correction (Mac). This Mac pre-training task was proposed to alleviate the inconsistency problem of pre-training to downstream tasks.

- ALBERT (A Lite BERT) [24]

ALBERT was proposed to improve the training and results of the BERT architecture using three different techniques: factorization of the embedding matrix, cross-layer parameter sharing, and inter-sentence coherence prediction.

- XLNet [25]

XLNet was proposed as a generalized autoregressive pre-training method that 1) enables learning of bidirectional contexts by maximizing the likelihood over all permutations of the factorization order; and 2) overcomes the limitations of BERT in neglecting dependencies between the masked positions. In addition, XLNet integrates the Transformer-XL mechanism into pretraining, which allows for the input of longer texts and reduces a pretrain-to-finetune discrepancy.

- ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements) [26]

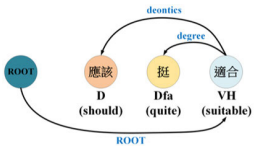
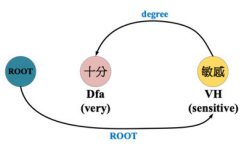
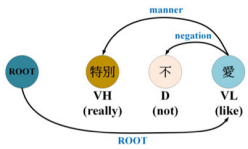
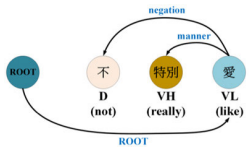
A new pre-training task called replaced token detection was proposed as an alternative to masking the input in BERT. ELECTRA consists of two parts: 1) generator: some tokens were replaced with plausible samples from a small generator network; and 2) discriminator: it predicts whether each token in the input was replaced by a generator or not.

Table 2 shows the results of sentiment intensity prediction on multi-word phrases. Our proposed Weighted-sum Tree GRU model outperforms the BERT [21], RoBERTa [22] and MacBERT [23], ALBERT [24], XLNet [25], and ELECTRA [26] models in both the valence and arousal dimensions, and equals the BERT result in Valence PCC. This indicates that the dependency paring can capture the modifier relationships between words, which helps enhance sentiment intensity prediction performance.

TABLE 2. Results of sentiment intensity prediction.

Model	Valence		Arousal		Number of Parameters (M)	Inference Time (millisecond)
	MAE	PCC	MAE	PCC		
BERT	0.487	0.936	0.484	0.866	110	660
RoBERTa	0.491	0.940	0.475	0.871	102	648
MacBERT	0.453	**0.948**	0.486	0.865	102	649
ALBERT	1.096	0.613	0.747	0.654	110	578
XLNet	0.535	0.929	0.564	0.814	117	680
ELECTRA	0.596	0.911	0.601	0.795	102	614
Weighted-sum Tree GRU	**0.392**	0.936	**0.399**	**0.915**	**59**	**278**

TABLE 3. Results of case study phrases.

Case Study Phrases	應該挺適合 (should be quite suitable)		十分敏感 (very sensitive)		特別不愛 (really not like)		不特別愛 (not really like)	
	V: 6.375	A: 4.333	V: 3.813	A: 6.375	V: 2.111	A: 6.714	V: 3.944	A: 4.929
Dependency Parsing Results								
BERT	6.647	4.050	3.546	5.819	4.003	5.648	6.435	4.497
RoBERTa	5.718	3.896	3.971	6.844	2.598	6.903	5.718	3.896
MacBERT	5.863	4.056	3.581	5.684	2.562	**6.735**	5.863	4.056
ALBERT	5.426	4.181	3.688	5.246	2.665	6.024	5.426	4.181
XLNet	4.857	4.793	3.161	5.973	2.751	5.969	4.857	4.548
ELECTRA	4.065	5.787	3.948	5.761	4.057	5.884	4.254	5.834
Weighted-sum Tree GRU	**6.032**	**4.301**	**3.826**	**6.225**	**2.432**	6.888	**3.943**	**4.975**

We also conducted a quantitative analysis to compare model size and inference time required. On a server using Nvidia GeForce RTX 2080 Ti GPUs with the same settings, the different transformer-based models require approximately 105M parameters and 640ms of inference time, while our proposed Weighted-sum Tree GRU model is relatively lightweight compared to the BERT-like transformer models, requiring only 43.8% of the number of parameters and 56.6% of the inference time, and does not require large amounts of data for pre-training.

In summary, our proposed Weighted-sum Tree GRU model is simple, but is effective and efficient in phrase-level sentiment intensity prediction due to the full use of linguistic dependency features for predicting sentiment intensity.

F. CASE STUDY

Table 3 shows the dependency parsing results of some phrases used for the case study and their valence-arousal rating predictions using the previously compared models. In the phrase "應該挺適合"(should be quite suitable), the modal

"應該" (should be) modifies the headword "適合" (suitable) with a deontic relation, which indicates the speaker's attitude towards whether an event is true or not. In addition, the degree adverb "挺" (quite) plays a semantic role to emphasize the statement. By obtaining these dependency features correctly, our Weighted-sum Tree GRU model predicts valence and arousal ratings respectively of 6.032 and 4.301, which are close to the human-annotated ratings of 6.375 and 4.333. In the phrase "十分敏感" (very sensitive), the degree adverb "十分" (very) is a behavioral relation used to modify the headword "敏感"(sensitive), which indicates how quickly the speaker reacts to an external stimulus. Our proposed Weighted-sum Tree GRU model respectively predicts valence and arousal ratings of 3.826 and 6.225, which are the nearest to the human-annotated ratings of 3.813 and 6.375. Moreover, we can identify two phrases "特別不愛" (really not like) and "不特別愛" (not really like) as having different meanings, despite having identical dependency modifiers. Our proposed Weighted-sum Tree GRU model can properly process the same modifiers in different orders, but with the same dependency structure. For the phrase"特別不愛" (human-annotated VA ratings are 2.111 and 6.714), our model predicts a valence of 2.432 and an arousal of 6.888. For the phrase "不特別愛" (VA ratings are 3.944 and 4.929), our model predicts very similar respective valence and arousal ratings of 3.943 and 4.975.

V. CONCLUSION

We propose a Weighted-sum Tree GRU network for phrase-level sentiment intensity prediction, making the following contributions:

(1) We develop a Chinese dependency parser based on the graph-based deep biaffine attention model to obtain dependency tree and relational information. Experimental results on the Sinica Treebank indicate that our graph-based model achieved a UAS of 92.9% and a LAS of 88.5%, which outperforms transition-based methods with identical findings for English dependency parsing.

(2) We propose a Weighted-sum Tree GRU model to include exploited dependency features for predicting Chinese phrase-level sentiment intensity in valence-arousal dimensions. Experimental results on the Chinese EmoBank indicate that our Weighted-sum Tree GRU model achieved an MAE of 0.392 and a PCC of 0.936 in the valence dimension and an MAE of 0.399 and a PCC of 0.915 in the arousal dimension, which outperforms several transformer-based models. Quantitative analysis also confirms that our model is relatively lightweight and efficient compared against BERT-like transformers, especially without the need of large amounts of data for pre-training.

Future work will exploit other semantic features and develop advanced models to further improve performance.

ACKNOWLEDGMENT

The authors sincerely appreciate ASUS TWCC for providing computing resources.

REFERENCES

[1] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.
[2] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion Measurement*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 201–237.
[3] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in *Proc. INDIACom*, 2016, pp. 452–455.
[4] S. Kiritchenko, S. Mohammad, and M. Salameh, "SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 42–51.
[5] F. Wang, Z. Zhang, and M. Lan, "ECNU at SemEval-2016 task 7: An enhanced supervised learning method for lexicon sentiment intensity ranking," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 491–496.
[6] A. Htait, S. Fournier, and P. Bellot, "LSIS at SemEval-2016 task 7: Using web search engines for English and Arabic unsupervised sentiment intensity prediction," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 469–473.
[7] L. Lenc, P. Král, and V. Rajtmajer, "UWB at SemEval-2016 task 7: Novel method for automatic sentiment intensity determination," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 481–485.
[8] E. Refaee and V. Rieser, "ILab-edinburgh at SemEval-2016 task 7: A hybrid approach for determining sentiment intensity of Arabic Twitter phrases," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 474–480.
[9] L.-C. Yu, L.-H. Lee, J. Wang, and K.-F. Wong, "IJCNLP-2017 task 2: Dimensional sentiment analysis for Chinese phrases," in *Proc. IJCNLP Shared Tasks*, 2017, pp. 9–16.
[10] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.
[11] C. Wu, F. Wu, Y. Huang, S. Wu, and Z. Yuan, "THU_NGN at IJCNLP-2017 task 2: Dimensional sentiment analysis for Chinese phrases with deep LSTM," in *Proc. IJCNLP Shared Tasks*, 2017, pp. 47–52.
[12] X. Zhou, J. Wang, X. Xie, C. Sun, and L. Si, "Alibaba at IJCNLP-2017 task 2: A boosted deep system for dimensional sentiment analysis of Chinese phrases," in *Proc. IJCNLP Shared Tasks*, 2017, pp. 100–104.
[13] P.-H. Li, W.-Y. Ma, and H.-Y. Wang, "CKIP at IJCNLP-2017 Task 2: Neural valence-arousal prediction for phrases," in *Proc. IJCNLP Shared Tasks*, 2017, pp. 89–94.
[14] S. Kiritchenko and S. M. Mohammad, "The effect of negators, modals, and degree adverbs on sentiment composition," 2017, *arXiv:1712.01794*.
[15] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Pipelined neural networks for phrase-level sentiment intensity prediction," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 447–458, Jul. 2020.
[16] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," 2016, *arXiv:1611.01734*.
[17] C.-R. Huang, F.-Y. Chen, K.-J. Chen, Z.-M. Gao, and K.-Y. Chen, "Sinica Treebank: Design criteria, annotation guidelines, and online interface," in *Proc. SIGHAN*, 2000, pp. 29–37.
[18] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 740–750.
[19] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-based dependency parsing with stack long short-term memory," 2015, *arXiv:1505.08075*.
[20] L.-H. Lee, J.-H. Li, and L.-C. Yu, "Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–18, Jul. 2022.
[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[23] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," 2020, *arXiv:2004.13922*.
[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.
[25] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[26] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.
[27] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Computing Attitude and Affect in Text: Theory and Applications*. Dordrecht, The Netherlands: Springer, 2006, pp. 1–10.
[28] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Comput. Intell.*, vol. 22, no. 2, pp. 110–125, 2006.
[29] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.
[30] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, Jan. 2012.
[31] J. Liu and S. Seneff, "Review sentiment scoring via a parse- and-paraphrase paradigm," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2009, pp. 161–169.
[32] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.
[33] J. Carrillo-de-Albornoz and L. Plaza, "An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 8, pp. 1618–1633, Aug. 2013.
[34] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, no. 1, 2014, pp. 216–225.
[35] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, no. 3. Cambridge, MA, USA: MIT Press, 2006.
[36] D. P. Solomatine and D. L. Shrestha, "AdaBoost.RT: A boosting algorithm for regression problems," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2004, pp. 1163–1168.
[37] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, 2013, pp. 1631–1642.
[38] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," 2015, *arXiv:1503.00075*.
[39] Y. Zhou, C. Liu, and Y. Pan, "Modelling sentence pairs with tree-structured attentive encoder," 2016, *arXiv:1610.02806*.
[40] L.-C. Yu, L.-H. Lee, S. Hao, J. Wang, Y. He, J. Hu, K. R. Lai, and X. Zhang, "Building Chinese affective resources in valence-arousal dimensions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 540–545.
[41] Y.-R. Wang and Y.-F. Liao, "A conditional random field-based traditional Chinese base phrase parser for SIGHAN bake-off 2012 evaluation," in *Proc. SIGHAN*, 2012, pp. 231–236.
[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[43] X. Ma, Z. Hu, J. Liu, N. Peng, G. Neubig, and E. Hovy, "Stack-pointer networks for dependency parsing," 2018, *arXiv:1805.01087*.

YU-CHIH DENG received the M.S. degree in communication engineering from the National Taipei University, Taiwan, in 2017. He is currently pursuing the doctoral degree with the Speech Processing Laboratory, Institute of Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan, under the guidance of Prof. Yih-Ru Wang and Sin-Horng Chen. His research interests include natural language processing and automatic speech recognition.

CHENG-YU TSAI received the B.S. and M.S. degrees in electrical engineering from the National Yang Ming Chiao Tung University, Taiwan, in 2017 and 2020, respectively. His research interests focus on natural language processing.

YIH-RU WANG received the B.S. and M.S. degrees from the Department of Communication Engineering, National Chiao Tung University, Taiwan, in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, National Chiao Tung University, in 1995. He served as an Associate Professor at the National Yang Ming Chiao Tung University, and retired in 2021. He is currently serves as a part-time Researcher. His general research interests include automatic speech recognition and natural language processing.

SIN-HORNG CHEN received the B.S. degree in communications engineering and the M.S. degree in electrical engineering from the National Chiao Tung University (NCTU), Taiwan, in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University, USA, in 1983. He was appointed as an Associate Professor and a Professor at the Department of Communications Engineering, NCTU, in 1983 and 1990, respectively. He served as the Dean at the ECE College, NCTU, and an Acting President of NCTU. He is currently the Chair Professor at the National Yang Ming Chiao Tung University. His major research interests include speech signal processing, particularly Mandarin speech recognition, text-to-speech, and speech prosody.

LUNG-HAO LEE (Member, IEEE) received the B.S. degree in statistics from the National Taipei University, Taiwan, in 2003, the M.S. degree in information management from Yuan Ze University, Taoyuan, Taiwan, in 2005, and the Ph.D. degree in computer science and information engineering from the National Taiwan University, in 2015. From 2015 to 2018, he was a Postdoctoral Fellow at the National Taiwan Normal University. He joined the Department of Electrical Engineering, National Central University, Taiwan, in 2018, as an Assistant Professor. He is currently an Associate Professor. His research interests include natural language processing, information retrieval and extraction, biomedical and health informatics, and artificial intelligence technologies.