Alibaba at IJCNLP-2017 Task 2: A Boosted Deep System for Dimensional Sentiment Analysis of Chinese Phrases

Xin Zhou and **Jian Wang** and **Xu Xie** and **Changlong Sun** and **Luo Si**
{eric.zx,eric.wj,Xu.Xie,luo.si}@alibaba-inc.com
changlong.scl@taobao.com

Abstract

This paper introduces Team Alibabas systems participating IJCNLP 2017 shared task No. 2 Dimensional Sentiment Analysis for Chinese Phrases (DSAP). The systems mainly utilize a multi-layer neural networks, with multiple features input such as word embedding, part-of-speech-tagging (POST), word clustering, prefix type, character embedding, cross sentiment input, and AdaBoost method for model training. For word level task our best run achieved MAE 0.545 (ranked 2nd), PCC 0.892 (ranked 2nd) in valence prediction and MAE 0.857 (ranked 1st), PCC 0.678 (ranked 2nd) in arousal prediction. For average performance of word and phrase task we achieved MAE 0.5355 (ranked 3rd), PCC 0.8965 (ranked 3rd) in valence prediction and MAE 0.661 (ranked 3rd), PCC 0.766 (ranked 2nd) in arousal prediction. In the final our submitted system achieved 2nd in mean rank.

1 Introduction

The task is to predict the affective states of a given (traditional) Chinese word in a continuous numerical value (score from 1 to 9) in the two-dimensional valence-arousal (V-A) space (Yu et al., 2016), indicating the degree from most negative to most positive for valence, and from most calm to most excited for arousal, which is the same as 2016s task. And in addition, predict the affective states of a given (traditional) Chinese phrase in the same V-A space. A human-tagged training data set containing 2802 words and 2250 phrases is used as the training set, another set of 750 words and 750 phrases is used as testing set. The result is measured by Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC) respectively.

This paper aims to present an introduction to Team Alibabas systems: data resources, feature engineering, model construction, and evaluation.

2 Chinese Corpus for Model Training

We have used the following text corpus with gratefulness for the openness of knowledge sharing.

1. Chinese Wikipedia dump with time stamp of 2017-07-20. There are over 1.3 million articles. Download link is at https://dumps.wikimedia.org/zhwiki/20170720/

2. Several forums dump (hot, boy-girl, movie, etc) from Taiwan online discussion board https://www.ptt.cc/bbs. There are around 70,000 articles. In addition to the main body, there are also 10 to 20 user comments in each article.

3. Liberty News Times articles from 2016-01-01 to 2017-08-15. We have used several subboards include Focus, Politics, Society, Local, Movie and Sports. There are around 100,000 articles.

All corpus is normalized to simplified Chinese characters before input into word embedding training. And after some evaluation only the first two sets of corpus are used to train the final model for V-A prediction.

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 100–104,
Taipei, Taiwan, November 27 – December 1, 2017.

3 Word Embedding

Although in the final model there are multiple features input, it is worthwhile describing word embedding in more details, because first, it has been a widely used representation of Chinese sentiment and semantic aspects recently, and second, it is also the input for other feature engineering. In our work, we have tried the following methods for word embedding in word V-A modeling.

- word2vec

  We have used open source python toolkit Gensim (Khosrovian et al., 2008) package. Skip-gram (Mikolov et al., 2013) methods are explored.

  These CWE models are trained with window size of 5, 15 iterations, 5 negative examples, minimum word count of 10, Skip-Gram (Mikolov et al., 2013) with starting learning rate of 0.025 , the output word vectors are of 300 dimensions.

- GloVe

  GloVe (Pennington et al., 2014) is an unsupervised learning algorithm for obtaining vector representations for words.

  The GloVe model is trained with window size of 8, minimum word count of 5 and maximum iteration of 20, the output word vectors are of 300 dimensions.

- Character-enhanced word embedding (CWE)

  Character-enhanced word embedding (Chen et al., 2015) (CWE) leverage composing characters to model the semantic meaning of word which shows effectiveness.

  The model setting is same as that of word2vec.

- cw2vec

  Cw2vec (Cao et al., 2017) proposes a stroke n-gram method for better handling of Chinese characters than Roman alphabets. An analogy to FastText (Bojanowski et al., 2016) that use sub-word information to enrich word embedding can help understand the method: in Chinese characters stroke n-gram is used as sub-word.

  The model setting is same as that of word2vec.

Different word embedding schemes are evaluated for both word and phrase level.

For phrase V-A modeling, word segmentation and average pooling are performed, and then word2vec is used to output word embedding.

4 Feature Engineering

In this section other more complex features are in addition to word embedding will be introduced.

For word level modeling

- CE: Average pooling of character embedding with 300 dimensions. The character embedding is trained with cw2vec as in section 3.

- CLU: Cluster feature of word. We use K-means to obtain 300 clusters with the word embedding trained with cw2vec as in section 3 and then represent the word cluster by one hot vector of 300 dimensions.

- POS: Part-of-speech-tagging (POST) of words containing verb, adverb, adjective, noun.

- VA: Words valence value used in arousal model training, and vice versa. The feature is represented by a one-dimension vector normalized to 1.

- POL: The polarity of word in NTUSD sentiment lexicon dictionary. The polarity of word is either positive of negative. The NTUSD sentiment dictionary is available at `http://academiasinicanlplab.github.io/`

For phrase level modeling

- TYPE: The prefix word of each phrase contains degree word (DEG),negative (NEG) word and modal word (MOD). As a result the prefix type is categorized into DEG / DEG-NEG /NEG-DEG / MOD-DEG / NEG / MOD-NEG / MOD. For example the type of "DEG-NEG" means the phase has a prefix with a degree word followed by a negative word. There are 7 prefix types, so this feature is represented by one hot vector of 7 dimensions.

- TAG: Word type feature. There are 4 types of words in the phrase - degree word (DEG), negative word (NEG), modal word (MOD),

sentiment word (SEN), so each word type can be represented by a one hot vector of 4 dimensions. Finally the TAG feature is represented by a concatenation of word type vector.

- CE: Same as word level

5 Model Construction

Inspired by (Du and Zhang, 2016) in submitted system, boosted neural network is used. Adaptive-Boosting (AdaBoosting) (Freund et al., 1996; Drucker, 1997) is used as boosting algorithm and there are 30 base regression models as most.

Neural network is used as base regression model with relu (Glorot et al., 2011) as activation function and Adam (Kingma and Ba, 2014) as its training algorithm and a constant learning rate of 0.001. For word V-A modeling the neural network is with 5 hidden layers and each layer is with 100/100/50/50/20 neurons. For phrase V-A modeling a one-layer neutral of 100 neurons in size network is used.

6 Evaluation

Evaluation is conducted locally by mean value of 5 rounds of 10 folds cross validation on training data, each round has constant and unique random seed.

6.1 Word level task

For word level we evaluate the performance of different word embedding methods and then we fix the word embedding type and evaluate different features.All manual features are converted to one hot vector or dense vector as a part of the input layer of neural network in addition to embedding features.

Word embedding comparison

Different embedding methods in section 3 are evaluated with boosted neural network, and we only report skip-gram schema in word2vec, CWE and cw2vec.

In Table 1, cw2vec outperforms all and CWE is slightly better than word2vec. Maybe we don't obtain the best hyper parameter so that GloVe gets worst performance.

Feature comparison

Different features in section 4 are evaluated in this part. WE denotes word embedding feature and other symbol are as listed in section 4.

Embeddings	Valence		Arousal	
	MAE	PCC	MAE	PCC
GloVe	0.605	0.809	1.241	0.618
word2vec	0.531	0.896	0.739	0.718
CWE	0.527	0.899	0.731	0.728
cw2vec	0.493	0.911	0.722	0.733

Table 1: Embedding comparison for word level

Features	Valence		Arousal	
	MAE	PCC	MAE	PCC
WE	0.493	0.911	0.722	0.733
WE+POS	0.491	0.913	0.723	0.734
WE+CE	0.460	0.924	0.683	0.768
WE+VA	0.495	0.908	0.723	0.726
WE+CE+POS	0.460	0.924	0.682	0.770
WE+CE+POL	0.437	0.932	0.677	0.773
WE+CE+POS+POL	0.435	0.933	0.675	0.773
WE+CE+POS+POL+CLU	0.413	0.938	0.567	0.840

Table 2: Feature comparison for word level

In Table 2, we can see CE (character embedding feature) and CLU(cluster feature) improve performance significantly. SEN (polarity feature) benefits valence prediction over arousal prediction, while POS feature and VA feature improve model performance slightly.

6.2 Phrase level task

For phrase level we also evaluate different pooling approaches besides the comparisons in word level task.

Embedding comparison

For phrase level experiment, phrase are segmented into words and use average pooling of word embedding to denote the phrase.

In Table 3 different from word level experiment word2vec achieves the best performance and GloVe under-performs other methods. Cw2vec is slightly better than CWE.

Embedding pooling comparison

As word2vec is fixed we evaluate the maximum pooling and average pooling. Table 4 shows average pooling is obviously better than maximum pooling as expected.

Features comparison

Now we have fixed the the embedding method

Embeddings	Valence		Arousal	
	MAE	PCC	MAE	PCC
GloVe	0.551	0.866	0.879	0.612
word2vec	0.462	0.937	0.434	0.883
CWE	0.479	0.929	0.439	0.88
cw2vec	0.475	0.934	0.438	0.881

Table 3: Embedding comparison for phrase level

Pooling	Valence		Arousal	
	MAE	PCC	MAE	PCC
average pooling	0.462	0.937	0.434	0.883
max pooling	0.590	0.897	0.610	0.851

Table 4: Pooling comparison for phrase level

Embeddings	Valence		Arousal	
	MAE	PCC	MAE	PCC
AWE	0.462	0.938	0.434	0.883
AWE+CE	0.463	0.937	0.434	0.884
AWE+POL	0.434	0.945	0.436	0.884
AWE+TAG	0.427	0.945	0.398	0.901
AWE+TYPE	0.416	0.948	0.396	0.903
AWE+TAG+POL	0.393	0.953	0.398	0.901
AWE+TYPE+POL	0.192	0.988	0.224	0.967

Table 5: Feature comparison for phrase level

to word2vec and average pooling is used. In this part AWE denotes average pooling of word embeddings, CE denotes average pooling of character embeddings. POL is a concatenation of one hot vector in words of phrase segmentation and padding 0 upto the longest length, so is TAG. Other features are described in section 4.

Table 5 presents the result that CE (character embedding feature) doesn't achieves positive result while TAG and TYPE and POL achieve extremely good performance. From the experiment we figure out the prefix type and polarity feature contains rich information for this task.

In the final submission for word level task, we use all features above. Run2 is an average boosted neural network applied with word embedding features on cw2vec and CWE while Run1 is generated on cw2vec alone. For phrase level task AWE, TYPE and POL are the best features used in Run1 and Run2. Run1 and Run2 use the same features and method with different random initial parameters for boosted neural network.

7 Conclusion

This system paper demonstrates Alibabas system for Dimensional Sentiment Analysis of Chinese Words and Phrases. We use boosted neural network as model for both word and phrase task. For word task cw2vec word embedding, average character embedding, cluster feature and polarity are identified as the best features. For phrase task word2vec word embedding, prefix type and polarity are identified as the best features. In the final test set we achieved MAE 0.545, PCC 0.892 in word valence estimation and MAE 0.857, PCC 0.678 in word arousal estimation and achieved MAE 0.526, PCC 0.901 in phrase valence estimation and MAE 0.465, PCC 0.854 in phrase arousal estimation. For average performance of word and phrase task we achieved MAE 0.5355(3rd), PCC 0.8965(3rd) in valence prediction and MAE 0.661(3rd), PCC 0.766(2nd) in arousal prediction. Our final submitted system achieved 2nd place in mean rank.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2017. Investigating chinese word embeddings based on stroke information.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *IJCAI*, pages 1236–1242.

Harris Drucker. 1997. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115.

Steven Du and Xi Zhang. 2016. Aicyber's system for ialp 2016 shared task: Character-enhanced word vectors and boosted neural networks. In *Asian Language Processing (IALP), 2016 International Conference on*, pages 161–163. IEEE.

Yoav Freund, Robert E Schapire, et al. 1996. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Keyvan Khosrovian, Dietmar Pfahl, and Vahid Garousi. 2008. Gensim 2.0: a customizable process simulation model for software process evaluation. *Making Globally Distributed Software Development a Success Story*, pages 294–306.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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 *HLT-NAACL*, pages 540–545.