abcbpc at SemEval-2021 Task 7: ERNIE-based Multi-task Model for Detecting and Rating Humor and Offense

Chao Pang, Xiaoran Fan, Weiyue Su, Xuyi Chen
Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang
Shikun Feng, Yu Sun
Baidu Inc., China
{pangchao04,fanxiaoran,suweiyue,chenxuyi}@baidu.com
{wangshuohuan,liujiaxiang,ouyangxuan}@baidu.com
{fengshikun01,sunyu02}@baidu.com

Abstract

This paper describes our system participated in Task 7 of SemEval-2021: Detecting and Rating Humor and Offense. The task is designed to detect and score humor and offense which are influenced by subjective factors. In order to obtain semantic information from a large amount of unlabeled data, we applied unsupervised pre-trained language models. By conducting research and experiments, we found that the ERNIE 2.0 and DeBERTa pre-trained models achieved impressive performance in various subtasks. Therefore, we applied the above pre-trained models to fine-tune the downstream neural network. In the process of fine-tuning the model, we adopted multi-task training strategy and ensemble learning method. Based on the above strategy and method, we achieved RMSE of 0.4959 for subtask 1b, and finally won the first place.

1 Introduction

Humor, as a highly subjective phenomenon, can be affected by various factors. Automatic humor recognition relies on annotated data to determine whether the text is humorous or not (Mihalcea and Strapparava, 2005). However, such a binary classification does not capture the level of humor, so assessing the level of humor is of great significance (Garimella et al., 2020). Since humor can be influenced by many factors, such as age, and may offend others. Based on such a situation, SemEval-2021 Task 7 focuses on linking humor and offense across different age groups (Meaney et al., 2021). But there are still many challenges to this task. For example, the dataset for the task is small and the texts are short, which does not allow for adequate training. To address these issues, we utilized unsupervised pre-trained language models and fine-tuned these models for specific downstream subtasks. After conducting research and extensive comparative experiments, the results show that ERNIE 2.0 (Sun et al., 2019b) and DeBERTa pre-trained models performed best on the subtasks. These large unsupervised language models were pre-trained on a large amount of unlabeled data to extract valuable lexical, syntactic, and semantic information from the corpus. Vector representations of text computed by these models are applied to fine-tune the downstream neural networks for the subtasks. The multi-task training and ensemble learning method significantly improve our model's performance.

The rest of the paper is organized as follows: Section 2 provides a brief description of the related work, and Section 3 describes our proposed approach in detail. In Section 4, the experiments are described in detail and the results are presented. Finally, we summarize the whole paper and discuss future research directions in Section 5.

2 Related Work

In the early research of humor and offense detection and evaluation, traditional machine learning methods and n-gram language model were mostly used.

Recent research has shown that unsupervised language pre-trained models using large amounts of unlabeled data have achieved state-of-the-art results in a large number of natural language processing tasks. For example, BERT (Devlin et al., 2018) is a model built based on Transformer Encoder, which is used for downstream tasks by pre-training on the masked language models task and the next sentence prediction task, and then for fine-tuning. Inspired by this approach, many pre-training language models have been proposed. For example, ALBERT (Lan et al., 2019) adopts Factorized Embedding Parameterization and Cross-layer parameter sharing strategies, and adds the sentence order prediction task, so that the model can greatly

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 286–289
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

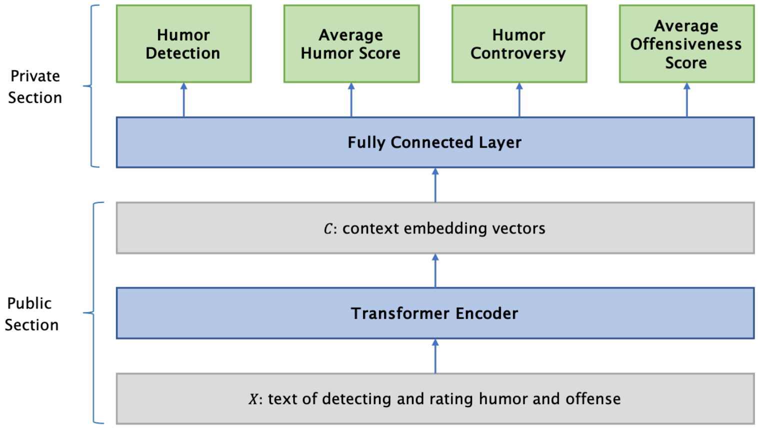

Figure 1: Architecture of our model for multi-task training. The public section is shared by all subtasks, while the private section is subtask-specific. First, text is transformed into subwords, and ERNIE 2.0 generates the corresponding contextual vector representation. Then different loss functions are set for each subtask to generate task-specific representations. This is eventually used for classification and regression subtasks in different scenarios.

reduce the number of parameters without lose accuracy compared to BERT. RoBERTa (Liu et al., 2019b) removes the next sentence prediction task and uses dynamic MASK, which is optimized for BERT. DeBERTa (He et al., 2020) improves the BERT and RoBERTa models by using two new techniques. The first one is using Disentangled attention mechanism and secondly, Enhanced mask decoder is used. MT-DNN (Liu et al., 2019a) combines Multi-task Learning and pre-trained models to improve the performance of various natural language processing tasks.

ERNIE 1.0 (Sun et al., 2019a) employs an entity-level and phrase-level mask, the extension of its training corpus and the use of multiple rounds of conversation to replace sentence pair classification further enhance the model's semantic representation capability. ERNIE 2.0 (Sun et al., 2019b) is an optimized version of ERNIE 1.0, which introduces a large number of pre-training tasks and continuously updates the pre-training model through multi-task learning to help the model learn lexical, syntactic and semantic representations efficiently. ERNIE 2.0 constructs three pre-training tasks, namely word-aware pre-training tasks, structure-aware pre-training tasks and semantic-aware pre-training tasks. The performance of the model is improved by constructing pre-training tasks from multiple perspectives. The ERNIE 2.0 model outperformed BERT and XLNet (Yang et al., 2019) almost across the board on the English task and achieved the best results on 7 GLUE tasks; on the Chinese task, the ERNIE 2.0 model outperformed BERT across the board on all 9 Chinese NLP tasks.

3 Our Approach

3.1 Multi-task training

To mitigate overfitting for specific tasks, we adopt a multi-task training strategy that combines pre-trained language model and multi-task training (as shown in Figure 1). Based on the above strategy, it makes the learned representation generalizable across tasks and improves the performance of various downstream subtasks.

The architecture of our model is shown in Figure 1, with the pre-trained model ERNIE 2.0 as the public section, which is used for generating semantic information common to downstream tasks. and the multi-task training as the private section, where individual subtasks are trained to produce task-specific representations by using different loss functions.

When we fine-tune our model, the input to the model is the data from all subtasks. Words from the text in different subtasks are first processed by tokenizer to generate subwords. After the subwords are transformed into tokens by the mapping of the lexicon, the tokenized sentences are stitched together with [CLS] and [SEP] as the input to the ERNIE 2.0 model to obtain the contextual vector representation corresponding to each token. Multi-task training is a fully-connected layer followed by the ERNIE 2.0 model, and the four downstream subtasks optimize the subtask-specific model by constructing different loss functions for gradient

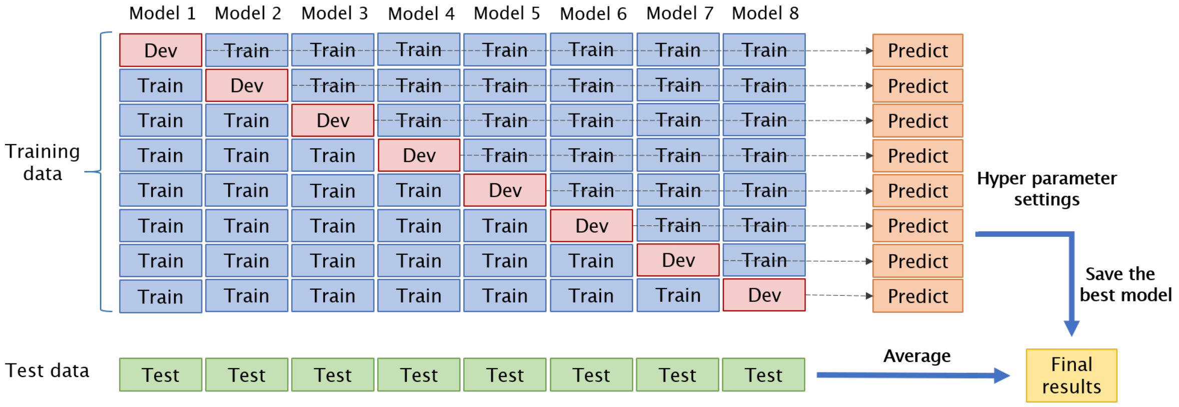

Figure 2: 8-fold cross-validation and ensemble. The training set is divided randomly for 8 times by setting different random seeds. In each division, the training set T is divided into 8 parts, of which 7 parts are respectively used as the training set and the remaining 1 part is used as the validation set. And finally the average of all saved best models predicted on the test set are the final results.

updating. For the classification task we first use the sigmod activation function to constrain the output between 0 and 1 before using the BCE loss function, and for the regression task we use the MSE loss function.

3.2 Ensemble

We adopt cross-validation for training as a way to improve the robustness of our model, as shown in Figure 2. We first divided the training set eight times by setting different random seeds. Therefore, 8 folds of data are generated, with 7000 training samples and 1000 validation samples in each fold. When fine-tune our model for each fold, the best model for each subtask at each fold of training is saved. For subtask 1a and subtask 1c, the evaluation metric is F1-Score, and for subtask 1b and subtask 2a the evaluation metric is RMSE. finally, we take the mean of all the best saved models after making predictions on the test set as the final results.

4 Experiment

4.1 Experimental details

All of our experiments were run on the Nvidia Tesla V100. In order to obtain more valuable information from the limited training data and to reduce overfitting to some extent, we adopt the multi-task training strategy and ensemble learning method.

For the training of the per-fold model, we choose the Adam optimizer, set the epoch to 10, and use early stopping strategies according to the performance on the validation set. Considering the small amount of data in the training set, we set a smaller learning rate for the ERNIE 2.0 model layer and a larger learning rate for the fully connected layer where the subtasks are trained together. Specifically, when fine-tuning our model, we adopt the grid search strategy with the learning rate ranging from 2e-5 to 5e-5 and the batch size ranging from 32 to 48. Besides, we set the learning rate as a linear function and use a warm-up strategy in the training phase. The ensemble approach we adopt is mainly based on the average prediction results. The specific methods are as follows: for the classification subtask, the prediction probabilities of all base models in each category are averaged, and then the category with the highest probability is taken as the prediction result; while for the regression subtask, the prediction values of all base models are averaged as the final prediction result.

4.2 Comparison experiments

In order to verify the effectiveness of our proposed multi-task training strategy based on the ERNIE 2.0 model, we set up two comparison experiments. They are described as follows:

(1) Comparison experiments of multi-task training together and single-task training separately based on ERNIE 2.0 model.

(2) Comparison experiments of single-task separate training based on ERNIE 2.0 and DeBERTa models.

4.3 Experimental Results

Table 1 summarizes the results on the validation set of all the models we tried based on the 8-fold cross-validation method. We can see that under

Models	Is_multi_task	Task 1a (F1-Score)	Task 1b (RMSE)	Task 1c (F1-Score)	Task 2a (RMSE)
DeBERTa$_{xlarge}$	False	0.9603	0.4492	0.6598	0.4817
ERNIE 2.0$_{xlarge}$	False	0.9656	0.4542	0.6388	0.4746
ERNIE 2.0$_{xlarge}$	True	0.9727	0.4475	0.6566	0.4722

Table 1: The results of different models under the 8-fold cross-validation method. The table describes the results of the DeBERTa and ERNIE 2.0 models on four subtasks in the case of single-task training separately and multi-task training together. For subtask 1a and subtask 1c, the evaluation metric is F1-Score, and for subtask 1b and subtask 2a the evaluation metric is RMSE.

the evaluation metrics of each subtask, The number of parameters in the ERNIE 2.0 model (425M) is less than the DeBERTa model (750M), but the ERNIE 2.0 pre-trained model performs better than the DeBERTa model on several subtasks. Besides, Compared to single-task training, the strategy of using multi-task training shows a significant improvement in performance. Moreover, The impact of ensemble learning method on improving model performance is significant.

5 Conclusion

In this paper, we proposed a multi-task training system based on ERNIE 2.0. We describe the architecture of the model and the training process in detail. Besides, we experimentally demonstrate that the strategy performs better with multi-task training compared to single-task training. Moreover, the ensemble learning method makes the model more robust. As a result, we have won the first place in a subtask for the competition of SemEval-2021 task 7. In our future work, we will further explore pre-trained language model and optimize the multi-task training.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. “judge me by my size (noun), do you?” YodaLib: A demographic-aware humor generation framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2814–2825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7, hahackathon, detecting and rating humor and offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019a. ERNIE: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019b. ERNIE 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.