SFC: FEW-SHOT TEXT CLASSIFICATION VIA SIMILARITY FUSED WITH CLASSIFICATION SYSTEM

Jingwen Huang

{hanmei613}@gmail.com

ABSTRACT

Building conversational chatbot system has become a popular solution to sharing the work of customer service under various business scenarios. A conversational chatbot needs to detect user's intent given a few words, which essentially equals to short-text classification problem in the field of Natural Language Processing. However, each time for a new service, the task-specific chatbot system often needs to perform well in few-shot setups due to lack of domain-specific data, which is still quite hard even if we use powerful pretrained model like Roberta. Therefore, in this paper, we propose SFC, a system fusion of both similarity model and classification model to overcome this challenge. Our main contributions are: 1) transfer learning and negative sampling based on sentence-pair are utilized as remedy for lack of data; 2) multi-task learning is involved to achieve faster training speed and better performance; 3) model ensembling of classification and similarity model guarantees inference speed while keeping high accuracy. Additionally, we also conduct extensive experiments on four public datasets in few-shot setup (i.e., with only 5 to 20 training data per class). The experimental results show that our system can steadily outperform several competitive baselines by 2 percent in average accuracy.

1. INTRODUCTION

Single-turn conversational chatbots are designed to transform existing tasks that rely on human agents, such as classifying customers' questions or queries to find the corresponding answer, into an automatic process based on intent classification model. Since user queries are usually much shorter than paragraphs or documents, the chatbot can actually be turned into a short-text Classification[1, 2, 3] task in Natural Language Processing. Moreover, at the initial stage of building chatbots, it's usually extremely hard to collect sufficient data for each class. In this way, building a single-turn conversational chatbot become a short-text classification[4] problem under few-shot setting[5].

Short texts[6] are usually more ambiguous in comparison with long texts since they don't contain enough contextual information, which poses a great challenge for classification[7]. In addition, the few-shot scenario[5] adds even more difficulties to the classification task since there is no enough information for the model to learn for each class. Comparing to non-pre-trained neural network such as convolutional neural networks (CNNs)[8, 9, 10] and long short term memory networks (LSTMs)[11, 12], the recently introduced pre-trained language models on large corpus like BERT[13] and RoBERTa[14] has been proven more powerful in solving many NLP tasks including short-text classification for deficient data[15]. Especially for few-shot scenarios, transfer learning based on pre-trained model tends to do more help to the negative effects brought by scarcity of training data. Due to the fact that RoBERTa is an improved version of BERT, we build our SFC chatbot system using RoBERTa as pre-trained context-dependent embeddings. To our knowledge, the most common approach is adding a softmax classifier to the top of RoBERTa(i.e., RoBERTa classifier), which turns the problem into a simple text classification task. However, despite the fact that RoBERTa has achieved amazing results in many Natural Language Processing tasks, we still believe it has much more potential under few-shot setting.

Therefore, in this paper we propose to involve similarity model (i.e., sentence-pair classifier) based on RoBERTa, which can score the semantic similarity between two sentences by multiple cross-attention mechanism[16]. A natural idea to enhance model performance in few-shot setting is to further pre-train RoBERTa with target domain data for transfer learning[17]. However, it's always not easy to find domain-specific data for certain service/product if we try to further pre-train a RoBERTa classifier directly. In comparison, it's relatively feasible to obtain dataset for semantically duplicate sentence pair identification task. That is to say, we can obtain a further pre-trained similarity model to identify the class with highest semantic similarity level to each user query. Afterwards, we can fine-tune the similarity model using sentence pairs sampled from target task dataset based on negative sampling strategy[18], which can provide us with more features to learn from limited amount of data.

We also apply multi-task learning[19, 20] in training process to enhance the training speed and performance. We set up two different objectives for training. The first one is the regular sentence-pair similarity score which help the model learn the semantic similarity between a query to each existing data point. The second target is to learn the similarity score between a query and all the data points of a certain class as a whole. Specific knowledge contained in these two tasks can be fully explored to obtain a faster training speed and higher accuracy.

Another contribution of this work is that we ensemble the similarity model with classification model at the inference step. In contrast with common model ensembling methods[21, 22], the classification model are used as an auxiliary model in our system to select promising sentence-pair candidates for similarity model. The experiment results on 4 short-text classification datasets in various few-shot settings show that our system can outperform single model baseline by at least 2 percent on average. Besides, we can also control the inference time for one single query to be within 0.5 seconds, which makes the system applicable in real-life single-turn chatbot scenario.

2. REFERENCES

[1] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 91–100.

[2] Rui Yan, Xian-bin Cao, and Kai Li, "Dynamic assembly classification algorithm for short text," *Acta Electronica Sinica*, vol. 37, no. 5, pp. 1019–1024, 2009.

[3] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou, "Short text understanding through lexical-semantic analysis," in *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 495–506.

[4] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 841–842.

[5] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou, "Diverse few-shot text classification with multiple metrics," *arXiv preprint arXiv:1805.07513*, 2018.

[6] Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie, "Short text classification: A survey," *Journal of multimedia*, vol. 9, no. 5, pp. 635, 2014.

[7] Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang, "Deep short text classification with knowledge powered attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6252–6259.

[8] Yoon Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[9] Xiang Zhang, Junbo Zhao, and Yann LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[10] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun, "Very deep convolutional networks for text classification," *arXiv preprint arXiv:1606.01781*, 2016.

[11] Amr Mousa and Björn Schuller, "Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 1023–1032.

[12] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, "Recurrent neural network for text classification with multitask learning," *arXiv preprint arXiv:1605.05101*, 2016.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[15] Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle, "Cost-sensitive bert for generalisable sentence classification with imbalanced data," *arXiv preprint arXiv:2003.11563*, 2020.

[16] Oren Barkan, Noam Razin, Itzik Malkiel, Ori Katz, Avi Caciularu, and Noam Koenigstein, "Scalable attentive sentence pair modeling via distilled sentence embedding.," in *AAAI*, 2020, pp. 3235–3242.

[17] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang, "How to fine-tune bert for text classification?," in *China National Conference on Chinese Computational Linguistics*. Springer, 2019, pp. 194–206.

[18] Robert Bamler and Stephan Mandt, "Extreme classification via adversarial softmax approximation," *arXiv preprint arXiv:2002.06298*, 2020.

[19] Rich Caruana, "Multitask learning: A knowledge-based source of inductive bias icml," *Google Scholar Google Scholar Digital Library Digital Library*, 1993.

[20] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[21] Leo Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.

[22] Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al., "Boosting the margin: A new explanation for the effectiveness of voting methods," *The annals of statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.