JOINT TRAINING OF CLASSIFICATION MODEL AND SIMILARITY MODEL FOR LOW-RESOURCE TEXT CLASSIFICATION IN CHATBOT

Jingwen Huang

{hanmei613}@gmail.com

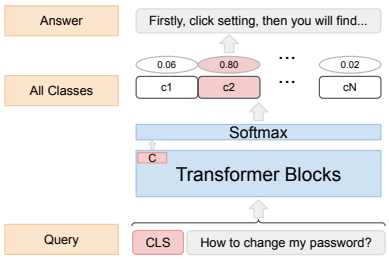
ABSTRACT

Building conversational chatbot system has become a popular solution to customer services 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. Moreover, 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 quite a challenge for single-model system such as text classification model system or sentence-pair semantic similarity model system under such a low-resource condition. Therefore, in this paper, we propose SFC, a 2-stage joint system with multi-task training technique for both text classification task and sentence-pair semantic similarity task to overcome this challenge. Furthermore, we additionally propose an improved version of SFC to allow text classification model and sentence-pair model be combined into a joint model organized in hierarchical structure. We also conduct extensive experiments on 4 public and 1 private 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 single-model baselines by 2 percent in average accuracy.

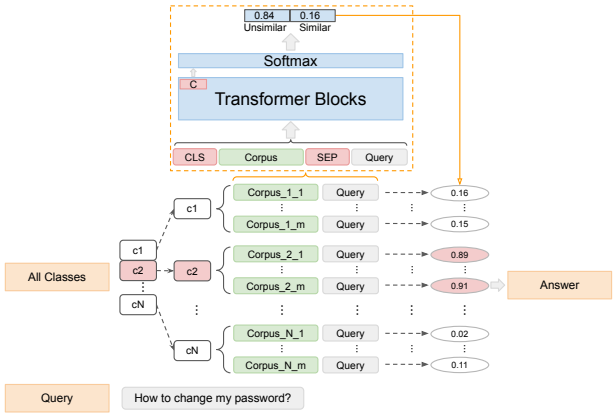
1. INTRODUCTION

Task-specific conversational chatbot[1] are designed to share the working pressure of human customer service agents who are responsible for solving customers' questions or queries about certain products or services. Regardless of whether the conversation is single-turn or multi-turn, the essential technical solution behind the chatbot is intent classification model which can identify the correct intent behind user's input text and find the corresponding answer.

Although task-specific conversational chatbot already has a wide application in various business and industries, it's still a quite challenging task due to its natural properties of low-resource. First, customers' utterance within a conversation is usually quite short and completed in a few words. Short text[2] are generally more ambiguous in comparison with long texts such as paragraphs or documents since they don't contain enough contextual information, which poses a great challenge[3] for classification task[4, 5, 6]. Second, at the initial stage of building a chatbot for specific task or service, it's often extremely hard to collect sufficient data samples for each class. The reason is that it's usually too expensive and time consuming to extract different ways of natural language expression for each intent class from history conversation log or even manually compose the corpus from nothing. Therefore, the key to building a task-specific chatbot with high performance becomes solving the challenge of short-text classification[7] problem under few-shot setting[8].



(a) text classification model structure based on pre-trained transformer



(b) sentence-pair model structure based on cross-attention mechanism

Fig. 1. Model Structure of 2 popular approach for building task-specific conversational chatbot

One of the most popular approach among existing work was text classification model based on neural network. Since the length of the text is quite short, many previous work[1] choose neural network such as convolutional neural networks (CNNs)[9, 10, 11] or long short term memory networks

(LSTMs)[12, 13] to accomplish the task of extracting semantic feature from limited amount of words. A common model structure is adding a softmax classifier to the top of the neural network. Afterwards, pre-trained language models on large corpus like BERT[14] and RoBERTa[15] has been proven more powerful in solving many NLP tasks including short-text classification[16]. Especially for few-shot scenarios[8], pre-trained model based on transformers[17], shown in Fig. 1(a), tends to do more help to reducing the negative effects brought by scarcity of training data.

Another popular approach among previous work was based on sentence-pair model. The motivation of this approach was started from the idea of Information Retrival (IR) based chatbot[18, 19]. Having a Q-A (Question-Answer) pairs dataset and user query Q, the IR based conversational system will look up in the Q-A dataset for the pair (Q', A') that best matches query Q through semantic analysis and returns A' as the answer to Q[20]. In this way, sentence-pair model pre-trained on large corpus of semantic similarity identification task can be applied for being used as a tool to identify the class with highest semantic similarity to customer's query. A common model structure of pre-trained[14] sentence-pair model is based on multiple cross-attention mechanism[21], shown in Fig. 1(b). Due to the fact that many experiment results have shown that RoBERTa[15] is an improved version of BERT, and has achieved amazing results on both text classification and sentence-pairs semantic similarity tasks, we choose RoBERTa as an important baseline approach in this paper.

Despite the success of these 2 approaches, they still have some limitations in task-specific chatbot scenario. As for the text classification model approach, it's quite hard to use data augmentation method to solve the data insufficiency challenge since the it's usually unfeasible to get large amount of domain-specific data when facing a new task. With respect to the sentence-pair model approach, though we can obtain large amount of data in semantically duplicate sentence pair identification domain for transfer learning[22], it's still quite hard to perform well on task-specific dataset because the training objective of sentence-pair model is semantic similarity, which is slightly different from the target of classification task. That is to say, there always exists some intent classes which cannot be distinguished by each other merely depending on semantic similarity. For example, we have 2 user query saying, A: What should I do if I want to change my password? B: What is the modification rule if I want to change my password? In this case, sentence A is semantically similar to B, since they both express the desire for changing password. However, it's still reasonable to classify them into 2 intent classes, since A is asking for the method for changing password, while B is asking for the rule to follow when creating a new password.

The above limitation motivates us to propose a joint system of both text classification task and sentence-pair semantic similarity task, named SFC here, shown in Fig. 2. Our

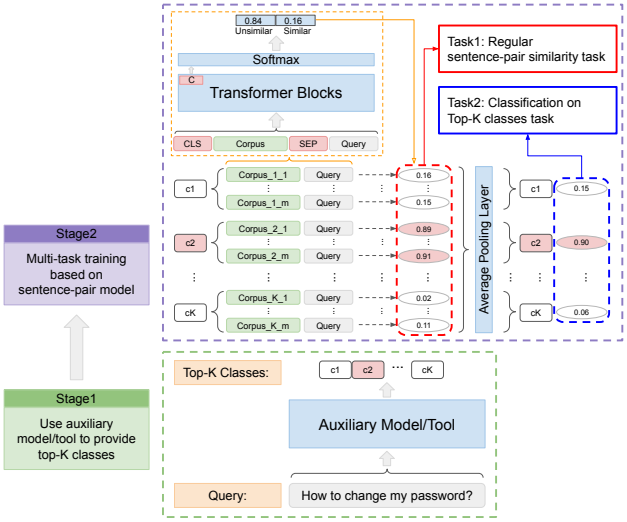


Fig. 2. Network Structure of 2-stage SFC: a joint system of multi-task for both text classification task and sentence-pair semantic similarity task.

goal is to utilize the advantage of the feasibility of transfer learning based on semantic similarity, while in the mean time, add the classification task's target into the training process of sentence-pair model for multi-task learning[23, 24]. To obtain such a joint system, we first start with preparation work, which is pre-training a sentence-pair model on external corpus for sentence pair duplication identification, and then find an auxiliary model or tool which can help us sample out top-K most related intent classes. This auxiliary model or tool can either be searching engine such as elasticsearch[25] or a text classification model trained on task-specific chatbot data. Afterwards, we further fine-tune the sentence-pair model with 2 training tasks: 1) the regular sentence pair similarity tasks. Here we use the text classification model as an auxiliary model in this paper to sample sentence pairs for training based on negative sampling strategy[26]. The reason is that a task-specific chatbot often has hundreds of intent classes, which forms too many sentence pairs for training since any two sentences from two different classes can form a sentence-pair of negative sample. Besides, according to Bamler[26], training on negative samples from most confusing wrong label can help model converge faster and obtain better performance. 2) the classification task on top-K (here K is a hyperparameter) candidate classes provided by our auxiliary model or tool. Here we use the average pooling of the representation given by sentence-pair model for sentence pairs formed by the user input query and the corpus sentences in each class as the feature. We stack the task-specific layers on the top of the shared-parameter sentence-pair model structure, which is a classic parameter sharing mechanism[24, 27, 28] for multi-task[22]. In this way, specific knowledge contained in these 2 tasks can be fully ex-

plored and deeply interact with each other to obtain better overall performance.

However, since the auxiliary tool in joint system mentioned above works in a separate stage from sentence-pair model, we find the the quality of sampled candidate classes might limit the overall performance of SFC, since the candidate classes for each user query is always fixed during the multi-task training process. This observation motivates us to further improve SFC into a joint model of sentence-pair model and text classification model in a hierarchical architecture by putting different tasks at different network layers[29, 30] and then training them together, shown in Fig 3. Here, the text classification model involved in joint training process replace the role of auxiliary tools. In this way, the sentence pairs selected from top-K classes becomes dynamic, which means the text classification model will also be further optimized to provide better top-K classes pooling result.

We summarize our contributions as follow:

1) We propose a novel joint system named SFC with multi-task training technique designed for task-specific conversational chatbot with low-resource, in which we make full utilization of the advantages brought by both the text classification task and sentence-pair semantic similarity task.

2) To further improve the system performance of SFC, we also propose an innovative joint model structure, in which the text classification model and sentence-pair model are fused into one single model for joint training.

3) Experiment results on 4 public and 1 private short-text classification datasets with respect to intent recognition tasks under conversational chatbot scenario all demonstrated that our proposed SFC joint system with multi-task, as well as the improved version of SFC, can both achieve remarkable improvement comparing to some powerful single-task and single-model baselines, especially under few-shot settings.

2. REFERENCES

[1] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562*, 2016.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke

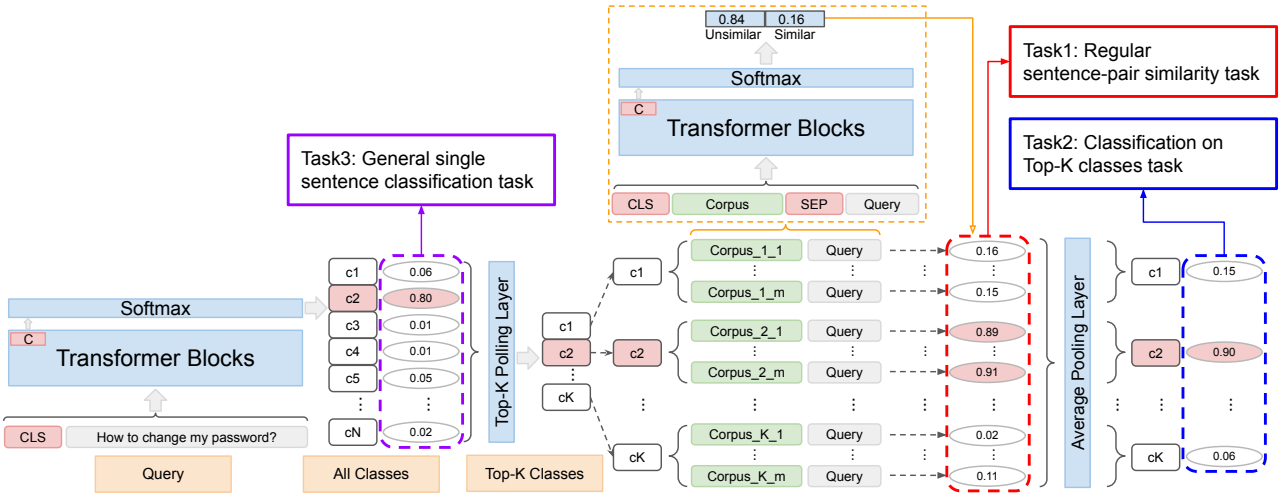


Fig. 3. **Network Structure of hierarchical SFC:** a joint model of sentence-pair model and text classification model organized in hierarchical structure by completing different tasks at different network layers.

Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

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

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] Sina Jafarpour, Christopher JC Burges, and Alan Ritter, "Filter, rank, and transfer the knowledge: Learning to chat," *Advances in Ranking*, vol. 10, pp. 2329–9290, 2010.

[19] Anton Leuski and David Traum, "Npceditor: Creating virtual human dialogue using information retrieval techniques," *Ai Magazine*, vol. 32, no. 2, pp. 42–56, 2011.

[20] Maali Mnasri, "Recent advances in conversational nlp: Towards the standardization of chatbot building," *arXiv preprint arXiv:1903.09025*, 2019.

[21] 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.

[22] 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.

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

[24] 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.

[25] Manda Sai Divya and Shiv Kumar Goyal, "Elasticsearch: An advanced and quick search technique to handle voluminous data," *Compusoft*, vol. 2, no. 6, pp. 171, 2013.

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

[27] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal, "Learning general purpose distributed sentence representations via large scale multitask learning," *arXiv preprint arXiv:1804.00079*, 2018.

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

[29] Anders Søgaard and Yoav Goldberg, "Deep multi-task learning with low level tasks supervised at lower layers," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 231–235.

[30] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher, "A joint many-task model: Growing a neural network for multiple nlp tasks," *arXiv preprint arXiv:1611.01587*, 2016.