Few-Shot Intent Classification by Gauging Entailment Relationship Between Utterance and Semantic Label

Jin Qu, Kazuma Hashimoto*, Wenhao Liu, Caiming Xiong, Yingbo Zhou
Salesforce Research, Palo Alto, CA, USA
{jqu,k.hashimoto,wenhao.liu,cxiong,yingbo.zhou}@salesforce.com

Abstract

Zhang et al. (2020) proposed to formulate few-shot intent classification as natural language inference (NLI) between query utterances and examples in the training set. The method is known as discriminative nearest neighbor classification or DNNC. Inspired by this work, we propose to simplify the NLI-style classification pipeline to be the entailment prediction on the utterance-semantic-label-pair (USLP). The semantic information in the labels can thus been infused into the classification process. Compared with DNNC, our proposed method is more efficient in both training and serving since it is based upon the entailment between query utterance and labels instead of all the training examples. The DNNC method requires more than one example per intent while the USLP approach does not have such constraint. In the 1-shot experiments on the CLINC150 (Larson et al., 2019) dataset, the USLP method outperforms traditional classification approach by >20 points (in-domain accuracy). We also find that longer and semantically meaningful labels tend to benefit model performance, however, the benefit shrinks as more training data is available.

1 Introduction

Many methods have been considered for few-shot intent classification. A simple but often effective approach is to simply generate more data through data augmentation. Wei and Zou, 2019 and Kumar et al., 2019 explored data augmentation at token- and feature-level to boost model performance. Meta-learning has also been studied extensively for few-shot learning. For instance, Induction Network (Geng et al., 2019) tried to learn general class representations via episode-based meta training and predict utterance labels based on the relation score between the query utterance and classes. Furthermore, large-scale pre-trained language models are often employed to mitigate the lack of annotated data for the target task. Schick and Schütze, 2021 leveraged pre-trained RoBERTa (Liu et al., 2019) and XLM-R (Conneau et al., 2020) to learn to generate task descriptions on small labeled datasets. They then use the trained models to produce descriptions and soft labels on large, task-specific unlabeled datasets and use them to train classifier. Although this approach has been proven to be effective, it requires extra unlabeled data and additional human supervision on description generation task. DNNC (Zhang et al., 2020) reformulates few-shot text classification as NLI-style pairwise comparison between training example and query utterance. However, DNNC requires at least two examples per intent for training and has to make $M \times$N (M: number of intents; N: number of training examples per intent) pairwise comparisons for each classification. Along the line of NLI-based classification, Yin et al. (2019) explored to leverage short semantic labels. However, this work is limited to zero-short setting and doesn't provide extensive analysis on how semantic information in labels affects model performance.

However, the DNNC work ignores one valuable and readily available supervision in the training data, the semantics in the labels. Our work is largely motivated by the hypothesis that semantic labels may carry valuable information about the intents and could benefit few-shot learning. There are prior works exploring how to leverage semantic labels in NLP tasks. For examples, Hou et al. (2020b) has proposed to improve the Prototypical Network (Snell et al., 2017) by directly embedding semantic labels; Hou et al. (2020a) has tried to use semantic information in labels for few-shot slot tagging. To our knowledge, however, there is no known work that has explored to leverage semantic labels for NLI-style intent classification. Neither has any work been done to study how model per-

*Work was done when the author was a full time research scientist at Salesforce Research, now works at Google

Proceedings of the Third Workshop on Natural Language Processing for Conversational AI, pages 8–15
November 10, 2021.

formance changes with regard to the interplay of data augmentation, different labeling, and number of training examples.

Based upon DNNC, our proposed method, utterance-semantic-label-pair (USLP), also leverages NLI-style classification. Instead of computing the entailment relationship between query sentence and the example sentences in the support set, we use the model to gauge the entailment relationship between query text and semantic labels. The semantic information in the labels can be perfectly infused into the classification process in this way. The pairwise entailment prediction is also reduced to M times per classification compared with the DNNC's $M\times$N. Figure 1 provides a few examples to illustrate the difference between USLP and DNNC.

In the following 1-shot experiments on the CLINC150 (Larson et al., 2019) dataset, we show that the USLP method outperforms the standard classification approach over 20 points with respect to in-domain accuracy. It is noteworthy that the predecessor of USLP, DNNC requires more than one example per intent for training. Although DNNC could do self-entailment training in 1-shot setting, our preliminary results show that the in-domain accuracy of multiple runs is extremely low (best result is below 20 from multiple runs). We also show that data augmentation, longer and more descriptive labeling, and NLI pre-training could boost model performance in few shot setting. However, as more training data is available, the efficacy of these performance boosters tends to shrink or even becomes negative. Our contributions can be summarized in two fold: 1, we proposed a new intent classification pipeline, USLP, and showed its effectiveness especially in 1-shot setting; 2, we studied how data augmentations, different labeling methods, and NLI pre-training might impact model performance in different few shot scenarios.

2 Method

2.1 Natural Language Inference

Natural Language Inference, or NLI, is a fundamental NLP task that aims to identify the relationship between a premise and a hypothesis. The relationship can be binary, (entailment and non-entailment) or ternary (entailment, contradiction, and neutral). Pre-trained transformer models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved promising results on NLI tasks. Here the NLI task is treated as a textual sequence classification problem, where the premise and hypothesis sentences are concatenated as $[CLS], premise, [SEP], hypothesis, [SEP]$ (depending on the tokenizer, the concatenated text might be slightly different) and fed into the model. The last hidden state of the $[CLS]$ token is commonly used for classification.

2.2 Utterance-Semantic-Label-Pair

The utterance-semantic-label-pair (USLP) approach builds on top of NLI framework as aforementioned. In USLP, utterances in training data are treated as premise while semantic labels are considered as hypothesis. We use binary entailment relationship for USLP, namely entailment and non-entailment. During training, an utterance-label pair is treated as a positive or entailment example if the label is the assigned intent for the utterance. Similarly, if the label is not the right intent label for the utterance, the pair is considered as a negative or non-entailment example. Although the USLP method does not necessarily require intent labels to have semantic meaning, detailed and semantically meaningful labels can benefit in-domain classification, which will be demonstrated in the following experiments. The DNNC (Zhang et al., 2020) method is also based upon NLI-style classification, the major difference is, it predicts the entailment relationship between the query utterance and examples in the training set. We provide Figure 1 to compare the two methods with more details.

For inference, we first generate all the possible query utterance and label pair and compute the entailment probability scores. The pair with the highest score has the predicted label for the utterance. To accommodate out-of-scope (OOS) prediction, we can either treat it same as an additional intent class like the other intent labels, or set up a threshold T, if the maximum entailment probability score is over T, we assign the corresponding label as the prediction, otherwise we assign OOS as the prediction for the query utterance.

3 Experiments

3.1 Datasets

3.1.1 General NLI corpus for pre-training

To unleash the full potential of transformer model on NLI task, we follow the data processing and training pipeline provided by Zhang et al. (2020) to combine three NLI corpus (SNLI (Bowman

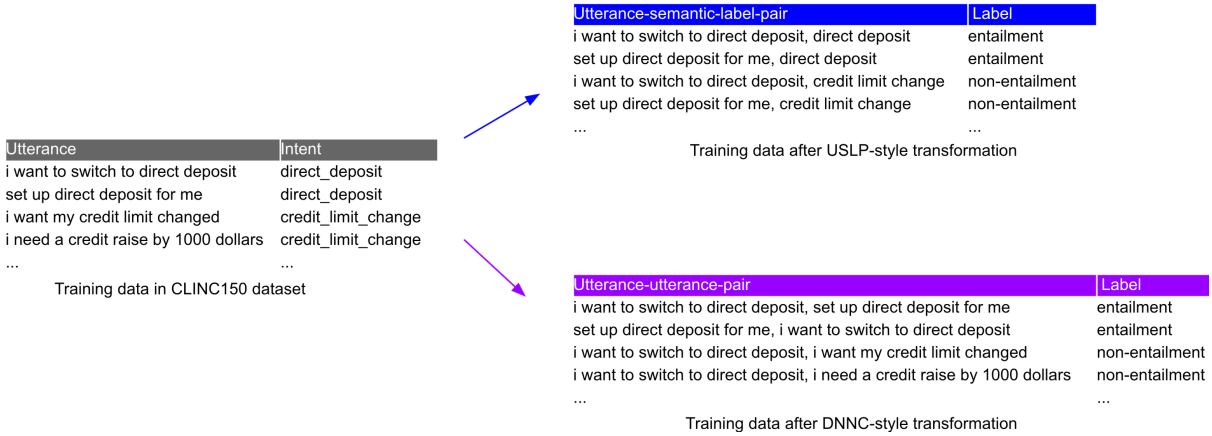

Figure 1: An illustration of training data in USLP and DNNC

et al., 2015), MNLI (Williams et al., 2018), and WNLI (Levesque, 2011)) from the GLUE benchmark (Wang et al., 2018) and use them for NLI pre-training.

3.1.2 CLINC150

CLINC150, introduced by Larson et al. (2019), is a multi-domain dataset for intent classification task. It has three dataset variants for in-domain and out-of-scope (OOS). We use the **small** dataset, which contains 150 intents, 50 examples/intent and 100 OOS examples for training. The original labeling has hyphen between each token in the label, we replace hyphen with empty space to format the label as short phrase. To simulate 1-, 5-, and 10-shot experiment, we randomly draw examples from the **small** dataset. We run each experiment five times with different seeds to capture the variations in random samplings. We remove dev set to simulate real few-shot scenario and use the original testing set for final results.

3.1.3 Schema-Guided Dialogue Dataset

SGD (Rastogi et al., 2019), or "Schema-Guided Dialogue Dataset", is a dataset about task-oriented dialogue. Its intent labels have detailed description, which is ideal for evaluating if detailed semantic labeling can help improve model performance. Since the original SGD dataset is not designed for few-shot intent classification, we went through a few data processing steps to customize the dataset for our use case. More details about the data processing steps can be found in Appendix B. We ended up with a subset of SGD dataset with 25 intents and 110 OOS utterances.

3.2 Data augmentation

We use the nlpaug library (Ma, 2019) for token-level data augmentation. In-domain utterances are augmented 4 times using random insertion, cBERT-based substitution, random swapping, and synonym replacement API. More details about the configurations can be found in Appendix A.

3.3 Training

We use the transformer library (4.5.1) (Wolf et al., 2020) by Huggingface for modeling. In NLI pre-training, we use the pre-trained roberta-base[1] model and follow the pre-training pipeline provided by Zhang et al. (2020). For downstream few-shot training, we use AdamW optimizer and linear scheduler, learning rate as 5e-5, epochs as 100, and train batch size as 128. We learnt this hyper-parameter set to be effective from our previous experiments with our in-house dataset. To simulate a real few shot setting, where dev set is often unavailable for hyper-parameter tuning and to demonstrate that the proposed method can be easily generalized into different datasets, we disregard all the dev sets in the following experiments and simply use the same hyper-parameter set without any further hyper-parameter tuning.

Balanced sampling Since the NLI reformulation of text classification results in much more negative examples than positive ones, we sample equal number of positive and negative examples for every batch to keep the model been exposed the balanced training examples. Furthermore, to prevent overfitting, each epoch iterates through all the positive examples while the negative examples are randomly

[1] https://huggingface.co/roberta-base

sampled to form batchs with positive examples. This data sampling strategy leads to better performance based upon previous empirical results on other in-house datasets. The previous DNNC (Zhang et al., 2020) work doesn't enforce balanced sampling, the positive and negative examples are mixed together and sampled randomly. We apply this sampling strategy to all the following experiments.

4 Results

4.1 Benchmark Results on CLINC150

USLP outperforms other methods by a large margin in 1-shot setting. Results from Table 1 show that USLP-T-A outperforms traditional classification approach by 20, 10, and 15 points in terms of in-domain accuracy, OOS recall, and OOS precision. The DNNC approch requires more than 1 example per class to start with, so it is out of the comparison. Compared with the 100-shot BERT-large results reported in Larson et al. (2019), the USLP-T-A achieves about 75% of the in-domain performance and has significantly higher OOS-recall score. Noticeably, within different USLP methods, the USLP-T has much better performance for in-domain accuracy (~20 points) and OOS-precision (>30 points) than USLP-O, but the USLP-O outperforms USLP-T by around 30 points for OOS-recall. One potential reason is the extremely unbalanced data; there are only one example per in-domain class, and in total we have 150 in-domain examples, but 100 examples for OOS. The USLP-O treats OOS as an extra class, but the OOS class has overwhelmingly more examples than other classes do, which could make the model favor OOS prediction. USLP-T approach, however, uses threshold to control in-domain and OOS prediction. Our experiments use 0.01 as the threshold, which tend to favor in-domain predictions and alleviates the extreme unbalance issue. Data augmentation can help improve in-domain classification and OOS-precision, but its impacts on OOS-recall and OOS-precision are opposite for USLP-T-A and USLP-O-A.

As we add more in-domain data into the experiments, in 5-shot and 10-shot experiments, we see the traditional classifier and DNNC in general perform better than USLP in terms of in-domain classification, but USLP has better and more balanced OOS-recall and OOS-precision scores. For example, in 10-shot experiments, CLS-T has the best in-domain accuracy, but it is unable to make OOS detection; DNNC has slightly better in-domain and OOS-precision result than USLP, but its OOS-recall is below that of USLP-T by around 30 points. Data augmentation seems to be more effective with USLP; it tends to hurt CLS and DNNC performance. Applying data augmentation on DNNC 10-shot training takes too much time (10+ hours on a single V-100 GPU), so we omit DNNC-A 10-shot experiment. Although the data augmentation continues to boost USLP in-domain performance for 5-shot and 10-shot experiments, it hurts OOS-recall. We believe that this is because the data augmentation will cause the model to be trained for more iterations due to fixed number of epochs and we sample 1/4 of batch size from OOS examples for every batch during training. As a result, the model is likely to overfit to the 100 OOS training examples.

4.2 The Role of Labeling Technique, Data Augmentation, and NLI pre-training

We use the SGD dataset to further study how relevant factors like labeling technique, data augmentation, and NLI pre-training on general corpus might impact USLP-T performance in different few-shot settings. Results are shown in Table 2.

Descriptive labeling can help improve USLP in-domain accuracy and OOS-precision. The SGD dataset provides intent labels as well as detailed descriptions for each label. To figure out the role of different labeling techniques in USLP-based intent classification, we set up three experiments with different labeling, 1) short labels, which are simply the original intent label. They are composed of either single words or short phrases and have limited semantic meaning; 2) long labels, which is the label description. Each description is usually a longer sentence than short labels and therefore can carry more semantic information; 3) symbolic labels. We convert labels into symbols like "0" and "1", which carry no semantic information. The results in Table 2 show that, long labels can effectively improve model performance. Especially at extreme low-resource scenario (1-shot), the long labels boost both in-domain accuracy and OOS-precision by 8+ points. Interestingly, long labels hurt model performance on OOS-recall. We hypothesize that long labels can boost model confidence on positive predictions resulting in producing higher prediction score favoring in-domain prediction.

	1-shot			5-shot			10-shot		
Method	**In-Acc**[3]	**OOS-R**[4]	**OOS-P**[5]	**In-Acc**	**OOS-R**	**OOS-P**	**In-Acc**	**OOS-R**	**OOS-P**
CLS-T[1]	51.56(1.31)	0.00(0.00)	NA	87.72(0.64)	0	NA	**92.52(0.41)**	0	NA
CLS-T-A[1]	51.44(1.61)	0.00(0.00)	NA	86.70(0.94)	0	NA	91.07(0.18)	0	NA
CLS-O[2]	43.41(3.17)	59.38(1.32)	48.20(4.82)	86.79(0.84)	42.98(2.36)	90.69(1.49)	91.95(0.50)	42.92(1.99)	95.05(1.01)
CLS-O-A[2]	45.31(2.28)	55.4(1.23)	51.03(5.00)	86.27(1.01)	43.18(1.81)	**92.84(1.75)**	91.18(0.40)	36.24(1.21)	**96.87(1.40)**
DNNC	NA	NA	NA	**88.49(1.00)**	61.46(6.10)	87.06(3.95)	91.21(0.61)	42.6(3.87)	92.69(1.39)
DNNC-A	NA	NA	NA	85.40(1.36)	20.30(10.33)	88.46(4.68)	NA	NA	NA
USLP-T	69.7(1.01)	62.62(3.07)	**66.92(3.64)**	83.96(1.45)	65.90(3.58)	84.32(3.26)	88.68(0.83)	**70.96(2.78)**	86.27(1.31)
USLP-T-A	**71.83(1.13)**	70.14(3.44)	65.18(5.08)	85.86(0.58)	65.47(2.81)	80.95(5.55)	90.29(0.44)	55.42(3.50)	89.23(2.30)
USLP-O	49.82(2.26)	**92.56(1.35)**	35.09(1.47)	79.28(1.11)	**67.06(2.34)**	72.61(2.45)	86.68(0.90)	56.70(3.81)	86.48(2.50)
USLP-O-A	66.84(1.05)	74.3(2.73)	55.67(3.40)	85.27(0.60)	54.18(3.83)	85.69(1.31)	90.22(0.59)	42.34(3.76)	94.51(1.40)
BL-100shot$_a$	96.9	40.3	NA						
BL-100shot$_b$	96.2	52.3	NA						

Table 1: CLINC150 few-shot benchmark results. [1] "CLS": traditional classifier using [CLS] token embedding for classification; "T": threshold, for all the experiments we use 0.01 as the threshold; "A": data augmentation; [2] "O": treating OOS as an additional class; BL-100shot$_a$ and BL-100shot$_b$ are based on bert-large model, reported by Larson et al. (2019). All other methods are based on roberta-base model. [3] "In-Acc", [4] "OOS-R", and [5] "OOS-P" stands for in-domain accuracy, OOS-recall, and OOS-precision respectively, numbers in the brackets represent standard deviation from multiple runs. DNNC requires >1 examples/intent for training and its 10-shot experiment with data augmentation takes >10 hours on a single V-100 GPU, so the corresponding experiments are skipped and results are shown as NA.

	1-shot			5-shot			10-shot		
Method	**In-Acc**	**OOS-R**	**OOS-P**	**In-Acc**	**OOS-R**	**OOS-P**	**In-Acc**	**OOS-R**	**OOS-P**
Short[1]	67.76(2.48)	**84.58(3.70)**	58.04(2.54)	85.54(1.18)	**74.53(6.63)**	89.84(2.33)	86.66(0.97)	**75.20(2.94)**	89.69(2.28)
Short-Aug[1]	69.68(3.51)	69.78(4.65)	60.35(4.34)	85.54(1.87)	67.96(5.17)	84.93(3.06)	85.46(1.00)	60.84(5.24)	84.85(2.27)
Long[2]	**76.24(2.47)**	77.2(3.96)	70.16(5.20)	**88.37(0.54)**	70.71(1.68)	**93.68(2.53)**	**88.59(0.61)**	74.80(4.11)	88.40(2.91)
Long-Aug[2]	74.22(1.14)	76.09(3.42)	**70.30(3.07)**	85.78(1.68)	74.00(5.03)	83.34(2.43)	87.14(0.83)	63.82(6.94)	86.44(2.05)
Symb[3]	4.58(1.02)	0	NA	7.94(1.71)	0	NA	58.90(8.18)	0	NA
Non-NLI[4]	64.64(3.74)	79.51(6.85)	62.96(1.21)	82.88(0.98)	62.00(3.55)	92.52(1.32)	88.26(0.51)	67.16(3.88)	**92.76(1.91)**

Table 2: USLP-T few-shot results on SGD dataset. [1] "short": original intent labels, which are either short phrases or single words; "Aug": data augmentation; [2] "long": detailed intent descriptions are used to replace short label to form utterance-label-pair; [3] "Symb": symbolic labels encoded as symbols like "0", "1", etc. They are converted from semantic labels; [4] "Non-NLI": the model is not fine-tuned on general NLI corpus.

Data augmentation is not always helpful. Quite different from the CLINC150 results, data augmentation fails to improve performance. In fact, data augmentation play a negative role in most experiments here. We tend to think that the effect of data augmentation is task-dependent, it might work well on some datasets but fail on other datasets. When developing few-shot applications with USLP, developers should be careful about applying data augmentation if no dev set is available.

NLI pre-training can boost performance in low-shot setting, but might have adverse effect when more training data is available. Our original hypothesis is that by exposing transformer model to NLI pre-training, the model can be more adapted into NLI related tasks and achieves better performance compared with the model without NLI pre-training. In 1-shot and 5-shot setting, we do observe that NLI pre-trained model can improve in-domain accuracy and OOS recall. But in 10-shot experiments, the NLI pre-trained model has weaker performance in terms of in-domain accuracy and OOS-precision.

5 Conclusion

We have created a new few-shot intent classification method, USLP, based upon NLI-style prediction. The USLP approach significantly outperforms traditional classification method by a large margin on 1-shot CLINC150 dataset and achieves about 75% of the 100-shot traditional classifier on in-domain classification with better OOS performance. This outstanding result indicates that the USLP approach can be an effective solution for developers who want to quickly build an intent classifier with extremely limited amount of training data. We have also found that detailed description can further boost USLP performance, but detailed labeling also requires labelers to have deeper understanding of each intent class and thus prolongs labeling process.

Acknowledgements

We would like to thank all the reviewers for their helpful comments. We would also like to thank Shashank Harinath, Shilpa Bhagavath, and Mridul Gupta for their insightful discussions on data augmentation methods.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3904–3913, Hong Kong, China. Association for Computational Linguistics.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020a. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online. Association for Computational Linguistics.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2020b. Few-shot Learning for Multi-label Intent Detection.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019. A Closer Look At Feature Space Data Augmentation For Few-Shot Intent Classification. pages 1–10.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Hector J Levesque. 2011. Levesque - The Winograd Schema Challenge. (1989).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1).

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 2017-Decem:4078–4088.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.

A Data Augmentation Methods

API name	action	model
WordEmbsAug	insert	GoogleNews-vectors-negative300
ContextualWord-EmbsAug	substitute	bert-base-uncased
RandomWordAug	swap	NA
SynonymAug	NA	ppdb-2.0-s-all

Table 3: Detailed configurations for nlpaug APIs.

B SGD Data Processing

We first extract utterances, intents, and detailed intent descriptions from the training set. The original labels formatted as tokens been concatenated together with the first letter capitalized, we introduce an empty space between each token. In the original dataset, the label set of the testing set does not fully overlap with the training set, so we keep the utterances with overlapped intents (25 intents) for in-domain and use the utterances with non-overlapped intents for OOS training (11 intents). Since our goal of using the SGD dataset is to explore how different labeling techniques might impact final results, we want to use the same training set to exclude the confounding factor of random training data sampling, so we sample 1-, 5-, 10-shot in-domain and 110 OOS (10 utterances/non-overlapped intent) utterances from the processed training set for all the SGD experiments. The original testing set has 11,105 utterances, which is expensive to run through for evaluation. So we sample 50 utterances per overlapped intents for in-domain testing set and 50 utterances per non-overlapped intents (9 non-overlapped intents) for OOS testing set, resulting in a testing set with 1,250 in-domain and 450 OOS utterances.