Effectiveness of Pre-training for Few-shot Intent Classification

Haode Zhang[1*] **Yuwei Zhang**[1*] **Li-Ming Zhan**[1]
Jiaxin Chen[1] **Guangyuan Shi**[1] **Xiao-Ming Wu**[1†] **Albert Y.S. Lam**[2]
Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.[1]
Fano Labs, Hong Kong S.A.R.[2]
haode.zhang@connect.polyu.hk, zhangyuwei.work@gmail.com
{lmzhan.zhan, jiax.chen, guang-yuan.shi}@connect.polyu.hk
xiao-ming.wu@polyu.edu.hk, albert@fano.ai

Abstract

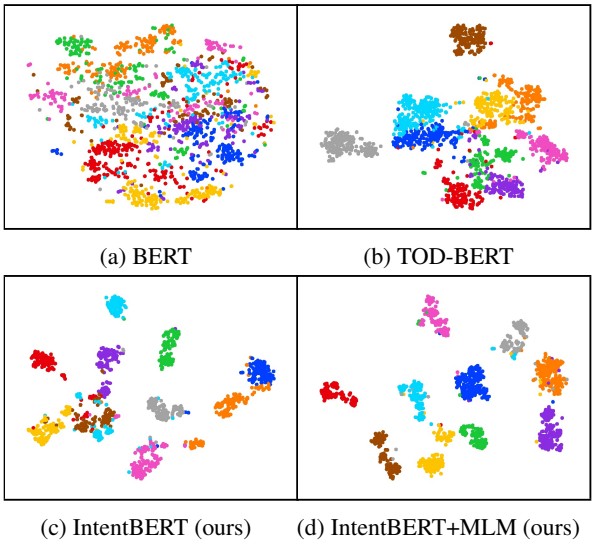
This paper investigates the effectiveness of pre-training for few-shot intent classification. While existing paradigms commonly further pre-train language models such as BERT on a vast amount of unlabeled corpus, we find it highly effective and efficient to simply fine-tune BERT with a small set of labeled utterances from public datasets. Specifically, fine-tuning BERT with roughly 1,000 labeled data yields a pre-trained model – IntentBERT, which can easily surpass the performance of existing pre-trained models for few-shot intent classification on novel domains with very different semantics. The high effectiveness of IntentBERT confirms the feasibility and practicality of few-shot intent detection, and its high generalization ability across different domains suggests that intent classification tasks may share a similar underlying structure, which can be efficiently learned from a small set of labeled data. The source code can be found at https://github.com/hdzhang-code/IntentBERT.

1 Introduction

Task-oriented dialogue systems have been widely deployed to a variety of sectors (Yan et al., 2017; Chen et al., 2017; Zhang et al., 2020c; Hosseini-Asl et al., 2020), ranging from shopping (Yan et al., 2017) to medical services (Arora et al., 2020a; Wei et al., 2018), to provide interactive experience. Training an accurate intent classifier is vital for the development of such task-oriented dialogue systems. However, an important issue is how to achieve this when only limited number of labeled instances are available, which is often the case at the early development stage.

To tackle few-shot intent detection, some recent attempts employ induction network (Geng et al., 2019), generation-based methods (Xia et al., 2020a,b), metric learning (Nguyen et al., 2020), or self-training (Dopierre et al., 2020). These works mainly focus on designing novel algorithms for representation learning and inference, which often comes with complicated models. Most recently, large-scale pre-trained language models such as BERT (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020) have shown great promise in many natural language understanding tasks (Wang et al., 2019), and there has been a surge of interest in fine-tuning the pre-trained language models for intent detection (Zhang et al., 2020a,b; Peng et al., 2020; Wu et al., 2020; Casanueva et al., 2020; Larson et al., 2019).

While fine-tuning pre-trained language models on large-scale annotated datasets has yielded significant improvements in many tasks including intent detection, it is laborious and expensive to construct large-scale annotated datasets in new application domains. Therefore, recent efforts have been dedicated to adapting pre-trained language models to a



(a) BERT (b) TOD-BERT

(c) IntentBERT (ours) (d) IntentBERT+MLM (ours)

Figure 1: Visualization of the embedding spaces with t-SNE. We randomly sample 10 classes and 500 data per class from BANKING77 (best viewed in color).

*Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1114–1120
November 7–11, 2021. ©2021 Association for Computational Linguistics

specific task such as intent detection by conducting continued pre-training (Gururangan et al., 2020; Gu et al., 2021) on a large unlabeled dialogue corpus with a specially designed optimization objective. Below we summarize the most related works in this line of research for few-shot intent detection.

- **CONVBERT** (Mehri et al., 2020) finetunes BERT on an unlabeled dialogue corpus consisting of nearly 700 million conversations.
- **TOD-BERT** (Wu et al., 2020) further pre-trains BERT on a task-oriented dialogue corpus of $100,000$ unlabeled samples with masked language modelling (MLM) and response contrastive objectives.
- **USE-ConveRT** (Henderson et al., 2020; Casanueva et al., 2020) investigates a dual encoder model trained with response selection tasks on 727 million input-response pairs.
- **DNNC** (Zhang et al., 2020a) pre-trains a language model with around 1 million annotated samples for natural language inference (NLI) and use the pre-trained model for intent detection.
- **WikiHowRoBERTa** (Zhang et al., 2020b) constructs some pre-training tasks based on the wikiHow database with $110,000$ articles.

While these methods have achieved impressive performance, they heavily rely on the existence of a large-scale corpus (Mehri et al., 2020) that is close in semantics to the target domain or consists of similar tasks for continued pre-training, which needs huge effort for data collection and comes at a high computational cost. More importantly, they completely ignore the "free lunch" – the publicly available, high-quality, manually-annotated intent detection benchmarks. For example, the dataset OOS (Larson et al., 2019) provides labeled utterances across 10 different domains. Hence, our study in this paper centers around the following research question:

- Is it possible to utilize publicly available datasets to pre-train an intent detection model that can *learn transferable task-specific knowledge to generalize across different domains*?

In this paper, we provide an affirmative answer to this question. We fine-tune BERT using a simple standard supervised training with approximately 1,000 labeled utterances from public datasets and obtain a pre-trained model, called IntentBERT. It can be directly applied for few-shot intent classification on a target domain that is drastically different from the pre-training data and significantly outperform existing pre-trained models, without further fine-tuning on target data (labeled or unlabeled). This simple "free-lunch" solution not only confirms the feasibility and practicality of few-shot intent detection, but also provides a ready-to-use well-performing model for practical use, saving the effort in algorithm design and data collection. Moreover, the high generalization ability of IntentBERT on cross-domain few-shot classification tasks, which are generally considered very difficult due to large domain gaps and the few data constraint, suggests that most intent detection tasks probably share a common underlying structure that could be learned from a small set of data.

Further, to leverage unlabeled data in the target domain, we design a joint pre-training scheme, which simultaneously optimizes the classification error on the source labeled data and the language modeling loss on the target unlabeled data. This joint-training scheme can learn better semantic representations and significantly outperforms existing two-stage pre-training methods (Gururangan et al., 2020). A visualization of the embedding spaces produced by strong baselines and our methods is provided in Fig. 1, which clearly demonstrates the superiority of our pre-trained models.

2 Methodology

We present a continued pre-training framework for intent classification based on the pre-trained language model BERT (Devlin et al., 2019).

Our pre-training method relies on the existence of a small labeled dataset $\mathcal{D}^{\text{labeled}}_{\text{source}} = \{(x_i, y_i)\}$, where y_i is the label of utterance x_i. Such data samples can be readily obtained from public intent detection datasets such as OOS (Larson et al., 2019) and HWU64 (Liu et al., 2021). As will be shown in the experiments, roughly $1,000$ examples from either OOS or HWU64 are enough for the pre-trained intent detection model to achieve a superior performance on drastically different target domains such as "Covid-19".

We further consider a scenario that unlabeled utterances $\mathcal{D}^{\text{unlabeled}}_{\text{target}} = \{x_i\}$ in the target domain are available, and propose a joint pre-training scheme

that is empirically proven to be highly effective.

2.1 Supervised Pre-training

Given $\mathcal{D}^{\text{labeled}}_{\text{source}} = \{(x_i, y_i)\}$ with N different classes, we employ a simple method to fine-tune BERT. Specifically, a linear layer is attached on top of BERT as the classifier, i.e.,

$$p(y|h_i) = \text{softmax}\left(\mathbf{W}h_i + \mathbf{b}\right) \in \mathbb{R}^N, \quad (1)$$

where $h_i \in \mathbb{R}^d$ is the feature representation of x_i given by the $[CLS]$ token, $\mathbf{W} \in \mathbb{R}^{N\times d}$ and $\mathbf{b} \in \mathbb{R}^N$ are parameters of the linear layer. The model parameters $\theta = \{\phi, \mathbf{W}, \mathbf{b}\}$, with ϕ being the parameters of BERT, are trained on $\mathcal{D}^{\text{labeled}}_{\text{source}}$ with a cross-entropy loss:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{ce}\left(\mathcal{D}^{\text{labeled}}_{\text{source}}; \theta\right). \quad (2)$$

After training, the fine-tuned BERT is expected to have learned general intent detection skills, and hence we call it IntentBERT.

2.2 Joint Pre-training

Given unlabeled target data $\mathcal{D}^{\text{unlabeled}}_{\text{target}}$, we can leverage it to further enhance our IntentBERT, by simultaneously optimizing a language modeling loss on $\mathcal{D}^{\text{unlabeled}}_{\text{target}}$ and the supervised loss in Eq. (2). The language modeling loss can help to learn semantic representations of the target domain while preventing overfitting to the source data.

Specifically, we use MLM as the language modeling loss, in which a proportion of input tokens are masked with the special token $[MASK]$ and the model is trained to retrieve the masked tokens. The joint training loss is formulated as:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{ce}(\mathcal{D}^{\text{labeled}}_{\text{source}}; \theta) + \lambda\mathcal{L}_{mlm}(\mathcal{D}^{\text{unlabeled}}_{\text{target}}; \theta), \quad (3)$$

where λ is a hyperparameter that balances the supervised loss and the unsupervised loss.

2.3 Few-shot Intent Classification

After pre-training, the parameters of IntentBERT are fixed, and it can be immediately used as a feature extractor for novel few-shot intent classification tasks. The classifier can be a parametric one such as logistic regression or a non-parametric one such as nearest neighbor. A parametric classifier will be trained with the few labeled examples provided in a task and make predictions on the unlabeled queries. As will be shown in the experiments, a simple linear classifier suffices to achieve very good performance, thanks to the effective utterance representations produced by IntentBERT.

3 Experiments

3.1 Experimental Setup

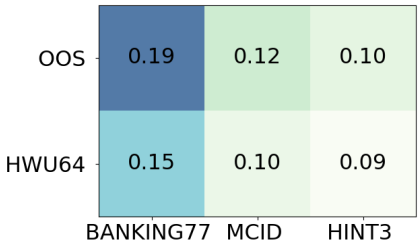


Figure 2: Vocabulary overlap.

Datasets. To train our IntentBERT, we continue to pre-train BERT on either of the two datasets, OOS (Larson et al., 2019)[1] and HWU64 (Liu et al., 2021), both of which contain multiple domains, providing rich resources to learn from[2]. For evaluation, we employ three datasets: **BANKING77** (Casanueva et al., 2020) is a fine-grained intent detection dataset focusing on "Banking"; **MCID** (Arora et al., 2020a) is a dataset for "Covid-19" chat bots; **HINT3** (Arora et al., 2020b) contains 3 domains, "Mattress Products Retail", "Fitness Supplements Retail" and "Online Gaming". Dataset statistics are summarized in Table 2.

Fig. 2 visualizes the vocabulary overlap between the source training data and target test data, which is calculated as the proportion of the shared words in the combined vocabulary of any two datasets after removing stop words. It is observed that the overlaps are quite small, indicating the existence of large semantic gaps.

Evaluation. The classification performance is evaluated by C-way K-shot tasks. For each task, We randomly sample C classes and K examples per class to train the classifier, and then we sample extra 5 examples per class as queries for evaluation. The accuracy is averaged over 500 such tasks.

Baselines. We compare IntentBERT to the following strong baselines. **BERT-Freeze** simply freeze the off-the-shelf BERT; **TOD-BERT** (Wu et al., 2020) further pre-trains BERT on a huge

[1]The domains "Banking" and "Credit Cards" are excluded because they are semantically close to the evaluation data.

[2]We have also experimented with the combination of both datasets but observed no better results.

Method	$\mathcal{D}^{\text{unlabeled}}_{\text{target}}$	BANKING77		MCID		HINT3	
		2-shot	10-shot	2-shot	10-shot	2-shot	10-shot
BERT-Freeze	✗	52.6 ±12.4	70.0 ±11.7	57.8 ±11.7	72.4 ±10.7	47.3 ±12.1	66.8 ±10.5
CONVBERT	✗	68.3 ±12.3	86.6 ±8.2	67.7 ±11.5	83.5 ±7.9	72.6 ±10.9	87.2 ±7.9
TOD-BERT	✗	77.7 ±7.4	89.4 ±5.1	64.1 ±9.0	77.7 ±11.0	68.9 ±11.7	83.5 ±8.6
USE-ConveRT¶	✗	–	85.2	–	–	–	–
DNNC	✗	67.5 ±15.4	89.8 ±7.5	56.2 ±16.7	80.0 ±9.9	64.1 ±14.8	87.9 ±8.1
WikiHowRoBERTa	✗	34.9 ±10.5	41.6 ±10.1	30.8 ±9.9	36.4 ±9.7	31.7 ±10.3	39.0 ±9.9
IntentBERT (HWU64) (ours)	✗	**78.4** ±10.6	**90.0** ±7.5	**74.5** ±11.9	**85.9** ±8.8	**77.9** ±10.6	**89.4** ±7.9
IntentBERT (OOS) (ours)	✗	**82.4** ±8.3	**91.8** ±4.2	**77.1** ±9.0	**88.1** ±5.9	**80.1** ±10.4	**90.2** ±7.4
IntentBERT (OOS)+MLM (ours)	✓	**88.9** ±9.0	**95.2** ±5.1	**86.3** ±9.8	**92.4** ±6.2	**87.1** ±9.8	**94.0** ±6.0

Table 1: Main results for 5-way tasks. ¶ stands for results from the original paper.

Dataset	#domain	#intent	#utterances
OOS	8	120	18000
HWU64	21	64	25716
BANKING77	1	77	13083
MCID	1	16	1745
HINT3	3	51	2011

Table 2: Dataset statistics.

amount of task-oriented conversations with MLM and response selection tasks; **CONVBERT** (Mehri et al., 2020) further pre-trains BERT on a large open-domain multi-turn dialogue corpus; **USE-ConveRT** (Henderson et al., 2020; Casanueva et al., 2020) is a fast embedding-based classifier pre-trained on an open-domain dialogue corpus by dialogue response selection tasks; **DNNC** (Zhang et al., 2020a) further pre-trains a BERT-based model on NLI tasks and then applies a similarity-based classifier for classification; **WikiHowRoBERTa** (Zhang et al., 2020b) further pre-trains RoBERTa (Liu et al., 2019) on fake intent detection data synthesized from wikiHow[3].

All the baselines (except BERT-Freeze) adopt a second pre-training stage, but with different objectives and on different corpus. In our experiments, all the baselines (except DNNC) use logistic regression as the classifier. For DNNC, we strictly follow the original implementation[4] to pre-train a BERT-style pairwise encoder to estimate the best matched training example for a query utterance.

Training details. We use $\text{BERT}_{\text{base}}$[5] (the base configuration with $d = 768$) as the encoder, Adam (Kingma and Ba, 2015) as the optimizer, and PyTorch library for implementation. The model is trained with Nvidia GeForce RTX 2080 Ti GPUs. For supervised pre-training, we use validation to control early-stop to prevent overfitting. Specifically, we use HWU64 for validation when pre-training with OOS and vice versa. The training is stopped if no improvement in accuracy is observed in 3 epochs. For joint pre-training, λ is set to 1. The number of training epochs is fixed to 10, since it is not prone to overfitting.

3.2 Main Results

The main results are provided in Table 1. First, IntentBERT (either pre-trained with OOS or HWU64) consistently outperforms all the baselines by a significant margin in most cases. Take the results of 5-way 2-shot classification on MCID for example, IntentBERT (OOS) outperforms the strongest baseline CONVBERT by an absolute margin of 9.4%, demonstrating the high effectiveness of our pre-training method. The cross-domain transferability of IntentBERT indicates that despite semantic domain gaps, most intent detection tasks probably share a similar underlying structure, which could be learned with a small set of labeled utterances. Second, IntentBERT (OOS) seems to be more effective than IntentBERT (HWU64), which may be due to the semantic diversity of the training corpus. Nevertheless, the small difference in performance between them shows that our pre-training method is not sensitive to the training corpus.

Finally, our proposed joint pre-training scheme (Section 2.2) achieves significant improvement over IntentBERT (up to 9.2% absolute margin), showing the high effectiveness of joint pre-training when target unlabeled data is accessible. Our joint pre-training scheme can also be applied to other language models such as GPT-2 (Radford et al., 2019) and ELMo (Peters et al., 2018), which is left as future work.

[3]https://www.wikihow.com/

[4]https://github.com/salesforce/DNNC-few-shot-intent

[5]https://github.com/huggingface/transformers

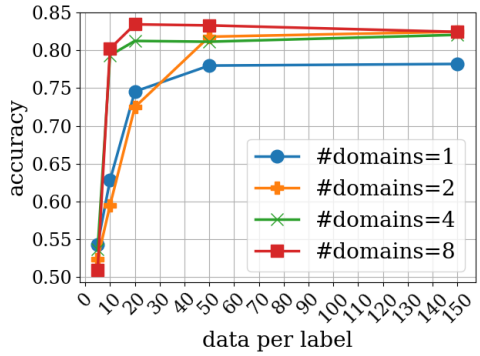


Figure 3: Effect of the amount of labeled data used for pre-training in the source domain (OOS). The results are evaluated on 5-way 2-shot tasks on BANKING77.

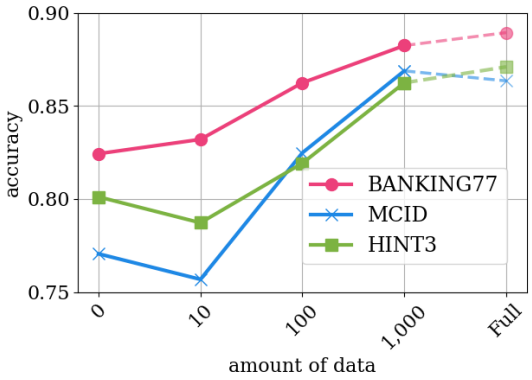


Figure 4: Effect of the amount of unlabeled data used for joint pre-training in the target domain. The results are evaluated on 5-way 2-shot tasks with OOS as the source dataset.

3.3 Analysis

Amount of labeled data for pre-training. We reduce the data used for pre-training in two dimensions: the number of domains and the number of samples per class. We randomly sample 1, 2, 4 and 8 domains for multiple times and report the averaged results in Fig. 3. It is found that the training data can be dramatically reduced without harming the performance. The model trained on 4 domains and 20 samples per class performs on par with that on 8 domains and 150 samples per class. In general, we only need around $1,000$ annotated utterances to train IntentBERT, which can be easily obtained in public datasets. This finding indicates that using small task-relevant data for pre-training may be a more effective and efficient fine-tuning paradigm.

Amount of unlabeled data for joint pre-training. We randomly sample a fraction of unlabeled utterances and re-run the joint training. As shown in Fig. 4, the accuracy keeps increasing when the number of unlabeled samples grows from 10 to $1,000$ and tends to saturate after reaching $1,000$. Surprisingly, $1,000$ utterances in BANKING77 can yield a comparable performance than the full dataset ($13,083$ utterances). Generally, it does not need much unlabeled data to reach a high accuracy.

Ablation study on joint pre-training. First, we investigate a two-stage pre-training scheme (Gururangan et al., 2020) where we use BERT or IntentBERT as initialization and perform MLM in the target domain (the top two rows in Table 3). It can be seen that they perform much worse than our joint pre-training scheme (the bottom row). Second, we use the source data instead of the target data for MLM in joint pre-training (the third row), and observe consistent performance drops, which shows the necessity of a domain-specific corpus.

Methods	BANK	MCID	HINT3
BERT→MLM(target)	80.5	63.0	72.0
IntentBERT→MLM(target)	82.0	75.9	77.9
IntentBERT+MLM(source)	84.1	75.9	78.5
IntentBERT+MLM(target)	**88.9**	**86.3**	**87.1**

Table 3: Ablation study on joint pre-training. BANK denotes BANKING77. → denotes moving to the next training stage. + denotes joint optimization of both loss functions. The data used for the experiment (either from "target" or "source") is shown in the brackets. The results are evaluated on 5-way 2-shot tasks with OOS as the source dataset.

4 Conclusion

We have proposed IntentBERT, a pre-trained model for few-shot intent classification, which is obtained by fine-tuning BERT on a small set of publicly available labeled utterances. We have shown that using small task-relevant data for fine-tuning is far more effective and efficient than current practice that fine-tunes on a large labeled or unlabeled dialogue corpus. This finding may have a wide implication for other tasks besides intent detection.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This research was supported by the grants of HK ITF UIM/377 and DaSAIL project P0030935.

References

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020a. Cross-lingual transfer learning for intent detection of covid-19 utterances.

Gaurav Arora, Chirag Jain, Manas Chaturvedi, and Krupal Modi. 2020b. HINT3: Raising the bar for intent detection in the wild. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 100–105, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

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

Thomas Dopierre, Christophe Gravier, Julien Subercaze, and Wilfried Logerais. 2020. Few-shot pseudo-labeling for intent detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4993–5003, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Jing Gu, Qingyang Wu, Chongruo Wu, Weiyan Shi, and Zhou Yu. 2021. PRAL: A tailored pre-training model for task-oriented dialog generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 305–313, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

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. *ArXiv preprint*, abs/1907.11692.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *ArXiv preprint*, abs/2009.13570.

Hoang Nguyen, Chenwei Zhang, Congying Xia, and Philip Yu. 2020. Dynamic semantic matching and aggregation network for few-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1209–1218, Online. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *ArXiv preprint*, abs/2005.05298.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Congying Xia, Caiming Xiong, Philip Yu, and Richard Socher. 2020a. Composed variational natural language generation for few-shot intents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3379–3388, Online. Association for Computational Linguistics.

Congying Xia, Caiming Xiong, Philip Yu, and Richard Socher. 2020b. Composed variational natural language generation for few-shot intents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3379–3388, Online. Association for Computational Linguistics.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4618–4626. AAAI Press.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020a. 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.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020b. Intent detection with WikiHow. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 328–333, Suzhou, China. Association for Computational Linguistics.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020c. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, pages 1–17.