Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning

Jian-Guo Zhang[1]*, **Trung Bui**[2], **Seunghyun Yoon**[2], **Xiang Chen**[2], **Zhiwei Liu**[1]
Congying Xia[1], **Quan Hung Tran**[2], **Walter Chang**[2], **Philip Yu**[1]
[1] University of Illinois at Chicago, Chicago, USA
[2]Adobe Research, San Jose, USA
{jzhan51, zliu213, cxia8, psyu}@uic.edu,
{bui, syoon, xiangche, qtran, wachang}@adobe.com

Abstract

In this work, we focus on a more challenging few-shot intent detection scenario where many intents are fine-grained and semantically similar. We present a simple yet effective few-shot intent detection schema via contrastive pre-training and fine-tuning. Specifically, we first conduct self-supervised contrastive pre-training on collected intent datasets, which implicitly learns to discriminate semantically similar utterances without using any labels. We then perform few-shot intent detection together with supervised contrastive learning, which explicitly pulls utterances from the same intent closer and pushes utterances across different intents farther. Experimental results show that our proposed method achieves state-of-the-art performance on three challenging intent detection datasets under 5-shot and 10-shot settings.

1 Introduction

Intent detection, aiming to identify intents from user utterances, is a key component in task-oriented dialog systems. In real systems such as Amazon Alexa, correctly identifying user intents is crucial for downstream tasks (Zhang et al., 2020b; Ham et al., 2020). A practical challenge is data scarcity as it is expensive to annotate enough examples for emerging intents, and how to accurately identify intents in few-shot learning has raised attention.

Existing methods address the few-shot intent detection tasks mainly from two perspectives: (1) data augmentation and (2) task-adaptive training with pre-trained models. For the first category, Zhang et al. (2020a) and Mehri et al. (2020b) propose a nearest neighbor classification schema with full use of the limited training examples in both training and inference stages. Xia et al. (2020b) and Peng et al. (2020) propose to generate utterances for emerging intents based on variational autoencoder (Kingma and Welling, 2013) and GPT-2 (Radford et al., 2019), respectively. For the second category, Casanueva et al. (2020) and Mehri et al. (2020a) conduct intent detection by leveraging related conversational pre-training models based on a few hundred million conversations. Meanwhile, they devise a task-adaptive training schema where the model is pre-trained on all relative intent datasets or the target intent datasets with mask language modeling.

However, previous methods such as data augmentation related models (Liu et al., 2021c) are inefficient for training and hard to scale to tasks with lots of intents. Moreover, these models do not tackle well the following scenarios: In real scenarios, the few-shot intent detection could be more challenging when there exist many fine-grained intents, especially semantically similar intents. For instance, BANKING77 (Casanueva et al., 2020) has a single domain with 77 intents, and CLINC150 (Larson et al., 2019) has ten domains with 150 intents. Many intents in the datasets are similar. Therefore, training models is rather challenging when there are only limited examples.

Inspired by the recent success of contrastive learning (He et al., 2020; Gunel et al., 2020; Chen et al., 2020; Radford et al., 2021; Liu et al., 2021a; Gao et al., 2021; Liu et al., 2021b), which aims to enhance discrimination abilities of models, this work proposes improving few-shot intent detection via Contrastive Pre-training and Fine-Tuning (CPFT). Intuitively, we first learn to implicitly discriminate semantically similar utterances via contrastive self-supervised pre-training on intent datasets without using any intent labels. We then jointly perform few-shot intent detection and supervised contrastive learning. The supervised contrastive learning helps the model explicitly learn to pull utterances from the same intent close and push utterances across different intents apart.

Our contributions are summarized as follows:

*Work done while the first author was an intern at Adobe Research.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1906–1912
November 7–11, 2021.

1) We design a simple yet effective few-shot intent detection schema via contrastive pre-training and fine-tuning. 2) Experimental results verify the state-of-the-art performance of CPFT on three challenging datasets under 5-shot and 10-shot settings.

2 Related Work

Since this work is related to few-shot intent detection and contrastive learning, we review recent work from both areas in this section.

The few-shot intent detection task typically includes three scenarios: (1) learn a intent detection model with only K examples for each intent (Zhang et al., 2020a; Mehri et al., 2020a; Casanueva et al., 2020); (3) learn to identify both in-domain and out-of-scope queries with only K examples for each intent (Zhang et al., 2020a, 2021; Xia et al., 2021b). (2) given a model trained on existing intents with all examples, learn to generalize the model to new intents with only K examples for each new intent (Xia et al., 2020a,b, 2021a).

In this work, we focus on the first scenario, and several methods have been proposed to tackle the challenge. Specifically, Zhang et al. (2020a) proposes a data augmentation schema, which pretrains a model on annotated pairs from natural language inference (NLI) datasets and designs the nearest neighbor classification schema to adopt the transfer learning and classify user intents. However, the training is expensive and hard to scale to tasks with hundreds of intents (Liu et al., 2020). Mehri et al. (2020b); Casanueva et al. (2020) propose the task-adaptive training, which leverages models pretrained from a few hundred million dialogues to tackle few-shot intent detection. It also includes an unsupervised mask language modeling loss on the target intent datasets and shows promising improvements.

Contrastive learning has shown superior performance on various domains, such as visual representation (He et al., 2020; Chen et al., 2020; Radford et al., 2021), graph representation (Qiu et al., 2020; You et al., 2020), and recommender systems (Liu et al., 2021b). Moreover, recent works also adopt contrastive learning in natural language processing tasks (Gunel et al., 2020; Liu et al., 2021a; Gao et al., 2021), which employs the contrastive learning to train the encoder. Specifically, (Gunel et al., 2020) designs a supervised contrastive learning loss for fine-tuning data. Gao et al. (2021) designs a simple contrastive learning framework through dropout and it shows state-of-the-art performance on unsupervised and full-shot supervised semantic textual similarity tasks. Liu et al. (2021a) designs self-supervised Mirror-BERT framework with two types of data augmentation: randomly erase or mask parts of the input texts; feature level augmentation through dropout.

Our work differs from them in several respects: Firstly, we specifically tackle the few-shot intent detection task rather than the general full-shot learning; Secondly, we design a schema and employ contrastive learning in both self-supervised pre-training and supervised fine-tuning stages.

3 CPFT Methodology

We consider a few-shot intent detection task that handles C user intents, where the task is to classify a user utterance u into one of the C classes. We set balanced K-shot learning for each intent (Zhang et al., 2020a; Casanueva et al., 2020), i.e., each intent only includes K examples in the training data. As such, there are in total $C \cdot K$ training examples.

In the following section, we first describe the self-supervised contrastive pre-training for utterance understanding before introducing the supervised fine-tuning for few-shot intent detection.

3.1 Self-supervised Pre-training

We retrieve the feature representation $\mathbf{h}_i$ for the i-th user utterance through an encoder model, which in this paper is BERT (Devlin et al., 2019), *i.e.*, $\mathbf{h}_i = \text{BERT}(u_i)$. We implicitly learn the sentence-level utterance understanding and discriminate semantically similar utterances through the self-supervised contrastive learning method (Wu et al., 2020b; Liu et al., 2021a; Gao et al., 2021):

$$\mathcal{L}_{\text{uns\_cl}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{h}_i, \bar{\mathbf{h}}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{h}_i, \bar{\mathbf{h}}_j)/\tau)}, \quad (1)$$

where N is the number of sentences in a batch. τ is a temperature parameter that controls the penalty to negative samples. $sim(\mathbf{h}_i, \bar{\mathbf{h}_i})$ denotes the cosine similarity between two input vectors $\mathbf{h}_i$ and $\bar{\mathbf{h}_i}$. $\bar{\mathbf{h}_i}$ represents the representation of sentence $\bar{u_i}$, where $\bar{u_i}$ is from the same sentence u_i but few (10%) tokens are randomly masked (Devlin et al., 2019). Specifically, we dynamically mask tokens during batch training (Wu et al., 2020a), i.e., a sentence has different masked positions across different training epochs, and we find it is beneficial

to the utterance understanding. The sentence u_i and $\bar{u_i}$ are inputted together to a single encoder during the batch training (Gao et al., 2021).

Besides the sentence-level enhancement, we also add the mask language modeling loss (Devlin et al., 2019; Wu et al., 2020a) to enhance the token-level utterance understanding:

$$\mathcal{L}_{\text{mlm}} = -\frac{1}{M}\sum_{m=1}^{M}\log P(x_m), \tag{2}$$

where $P(x_m)$ denotes the predicted probability of a masked token x_m over the total vocabulary, and M is the number of masked tokens in each batch.

Our total loss for each batch is $\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{uns\_cl}} + \lambda\mathcal{L}_{\text{mlm}}$, where λ is a weight hyper-parameter.

3.2 Supervised Fine-tuning

Through self-supervised learning in the first stage, the model efficiently utilizes many unlabeled user utterances. The model is given very limited examples in the second stage, such as 5 and 10 examples for each intent. To better understanding user intents, especially when intents are similar to each other, we utilize a supervised contrastive learning method (Gunel et al., 2020) and train it together with an intent classification loss. We treat two utterances from the same class as a positive pair and the two utterances across different classes as a negative pair for contrastive learning. Unlike the previous work, the utterance and itself could also be a positive pair as we input them together to the single encoder. Their feature representations are different due to the dropout of BERT. The corresponding loss is shown as the following:

$$\mathcal{L}_{\text{s\_cl}} = -\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{1}_{y_i=y_j}\log\frac{\exp(\text{sim}(\mathbf{h}_i,\mathbf{h}_j)/\tau)}{\sum_{n=1}^{N}\exp(\text{sim}(\mathbf{h}_i,\mathbf{h}_n)/\tau)}, \tag{3}$$

where T is the number of pairs from the same classes in the batch.

Next is the intent classification loss:

$$\mathcal{L}_{\text{intent}} = -\frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N}\log P(C_j|u_i), \tag{4}$$

where $P(C_j|u_i)$ is the predicted probability of the i-th sentence to be the j-th intent class.

We jointly train the two losses together at each batch: $\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{s\_cl}} + \lambda'\mathcal{L}_{\text{intent}}$, where λ' is a weight hyper-parameter.

Name	# Utterance	# Intent	# Domain
CLINC150 (Larson et al., 2019)	18200	150	10
BANKING77 (Casanueva et al., 2020)	10162	77	1
HWU64 (Liu et al., 2019)	10030	64	21
TOP (Gupta et al., 2018)	35741	25	2
SNIPS (Coucke et al., 2018)	9888	5	-
ATIS (Tur et al., 2010)	4978	21	-

Table 1: Data statistics for intent detection datasets.

4 Experimental Settings

4.1 Datasets

Pre-training Datasets We collected six public datasets consisting of different user intents. The dataset statistics are shown in Table 1.[1] For fair comparisons, we exclude their test sets during the pre-training phase, which is different from previous work (Mehri et al., 2020a,b), where they use the whole datasets. We also remove utterances with less than five tokens, and there are 80,782 training utterances in total. We conduct self-supervised pre-training on the collected utterances without using labels.

Evaluation Datasets To better study the more challenging fine-grained few-shot intent detection problem and compare with recent state-of-the-art baselines, we pick up three challenging intent detection datasets for evaluation, *i.e.*, CLINC150 (Larson et al., 2019), BANKING77 (Casanueva et al., 2020) and HWU64 (Liu et al., 2019). CLINC150 contains 23,700 utterances across ten different domains, and there are in total 150 intents. BANKING77 contains 13,083 utterances with a single banking domain and 77 intents. HWU64 includes 25,716 utterances with 64 intents spanning 21 domains. We follow the setup of Mehri et al. (2020a), where a small portion of the training set is separated as a validation set, and the test set is unchanged. Following previous work, we repeat our few-shot learning model training five times and report the average accuracy.

4.2 Model Training and Baselines

We utilize RoBERTa with `base` configuration, *i.e.*, `roberta-base` as the BERT encoder. We pre-train the combined intent datasets without test sets in the contrastive pre-training stage for 15 epochs, where we set the batch size to 64, τ to 0.1, and λ to 1.0. The pre-training phase takes around 2.5 hours on a single NVIDIA Tesla V100 GPU with 32GB memory. We fine-tune the model under 5-shot

[1] https://github.com/jianguoz/Few-Shot-Intent-Detection

Model	CLINC150		BANKING77		HWU64	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
RoBERTa+Classifier (Zhang et al., 2020a)	87.99	91.55	74.04	84.27	75.56	82.90
USE (Casanueva et al., 2020)	87.82	90.85	76.29	84.23	77.79	83.75
CONVERT (Casanueva et al., 2020)	89.22	92.62	75.32	83.32	76.95	82.65
USE+CONVERT (Casanueva et al., 2020)	90.49	93.26	77.75	85.19	80.01	85.83
CONVBERT (Mehri et al., 2020a)	-	92.10	-	83.63	-	83.77
CONVBERT + MLM (Mehri et al., 2020a)	-	92.75	-	83.99	-	84.52
CONVBERT + Combined (Mehri et al., 2020b)	-	93.97	-	85.95	-	86.28
DNNC (Zhang et al., 2020a)	91.02	93.76	80.40	86.71	80.46	84.72
CPFT	**92.34**	**94.18**	**80.86**	**87.20**	**82.03**	**87.13**

Table 2: Testing accuracy ($\times 100\%$) on three datasets under 5-shot and 10-shot settings.

(5 training examples per intent) and 10-shot settings (10 training examples per intent). We set the batch size to 16, and do hyper-parameters search for $\tau \in \{0.1, 0.3, 0.5\}$ and $\lambda' \in \{0.01, 0.03, 0.05\}$; the fine-tuning takes five minutes for each run with 30 epochs. We apply label smoothing to the intent classification loss, following Zhang et al. (2020a).

Baselines We compare with six strong models. 1, RoBERTa+Classifier (Zhang et al., 2020a): it is a RoBERTa-based classification model. 2, USE Yang et al. (2020): it is the large multilingual model pre-trained on 16 languages. 3, CONVERT (Casanueva et al., 2020): it is an intent detection model with dual encoders, and the dual encoder models are pre-trained on 654 million (input, response) pairs from Reddit. 4, CONVEBERT (Mehri et al., 2020a): it fine-tunes BERT on a large open-domain dialogue corpus with 700 million conversations. 5, CONVEBERT+Comined (Mehri et al., 2020b): it is an intent detection model based on CONVEBERT, with example-driven training based on similarity matching and observers for transformer attentions. It also conducts task-adaptive self-supervised learning with mask language modeling (MLM) on the intent detection datasets. `Combine` represents the best `MLM+Example+Observers` setting in the referenced paper. 6, DNNC (Zhang et al., 2020a): it is a discriminative nearest-neighbor model which finds the best-matched example from the training set through similarity matching. The model conducts data augmentation during training and boosts performance by pre-training on three natural language inference tasks.

5 Experimental Results

We show the overall comparisons on three datasets in Table 2. The proposed CPFT method achieves the best performance across all datasets under both the 5-shot and 10-shot settings. Specifically, CPFT outperforms DNNC by 1.32% and 1.57% on CLINC150 and HWU64 under the 5-shot setting, respectively. It also improves DNNC by 2.41% on HWU64 under the 10-shot setting. Our variances are also lower when compared with DNNC: Ours vs. DNNC: 0.39 vs. 0.57 and 0.18 vs. 0.42 on CLINC150; 0.20 vs. 0.88 and 0.48 vs. 0.21 on BANKING77; 0.51 vs. 1.00 and 0.25 vs. 0.38 on HWU64 under 5-shot and 10-shot settings, respectively. The improvements indicate that our proposed method has a better ability to discriminate semantically similar intents than the strong discriminate nearest-neighbor model with data augmentation. Moreover, the DNNC training is expensive, as when training models on a single NVIDIA Tesla V100 GPU with 32GB memory, DNNC takes more than 3 hours for 10-shot learning on CLINC150, and it needs to retrain the model for every new setting. CPFT only needs 2.5 hours for one-time pre-training, and the fine-tuning only takes five minutes for each new setting. Compared with CONVBERT+MLM, which does a self-supervised pre-training with MLM on the intent detection datasets, CPFT improves the performance by 1.43%, 3.21%, and 2.61% on CLINC150, BANKING77, and HWU64 under 10-shot setting, respectively. CPFL also outperforms CONVBERT+Combined, which further adds examples-driven training and specific transformer attention design. We contribute the performance improvements to contrastive learning, which help the model discriminate semantically similar intents.

6 Ablation Study and Analysis

Is the schema with both stages necessary? We conduct ablation study to investigate the effects

Model	CLINC150		BANKING77		HWU64	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
CPFT	92.34	94.18	80.86	87.20	82.03	87.13
w/o Contrastive pre-training	-4.15	-2.63	-4.11	-2.37	-6.01	-4.17
w/o Supervised contrastive learning	-0.56	-0.32	-2.06	-0.88	-1.14	-0.27
w/o Contrastive pre-training + w/o Supervised contrastive learning	-4.35	-2.69	-6.82	-2.93	-6.47	-4.23

Table 3: Testing accuracy ($\times 100\%$) of CPFT with variants on three datasets under 5-shot and 10-shot settings.

of self-supervised contrastive pre-training and supervised contrastive fine-tuning. Table 3 shows the testing results of CPFT with model variants on three datasets. Experimental results indicate that both stages are necessary to achieve the best performance. The self-supervised contrastive pre-training on the first stage is essential as the performance drops significantly on all datasets. We hypothesize that contrastive pre-training on the intent datasets without using labels benefits the discrimination of semantically similar utterances. Additionally, the performance also drops if without supervised contrastive learning during the few-shot fine-tuning stage. Specifically, it drops by 2% on BANKING77 under the 5-shot setting; the reason is that BANKING77 is a single domain dataset with many similar intents, where supervised contrastive learning can explicitly discriminate semantically similar intents with very limited training examples. We also jointly train the first and second stages together, and compared with the proposed CPFT schema, we observe minimal improvements. The joint training is also costly as it requires retraining the model every time for new settings.

Is contrastive pre-training beneficial to the target intent dataset? Additionally, we study whether contrastive pre-training can benefit the intent detection when excluding the target datasets. Specifically, we pre-train the model on the datasets except for the HWU64 dataset on the first stage and do few-shot learning on HWU64 during the second stage. Compared to the model without contrastive pre-training on the first stage, the performances are improved by 1.98% and 1.21% under 5-shot and 10-shot settings, respectively. The improvements indicate that the contrastive pre-training is helpful to transfer knowledge to new datasets. However, there are still performance drops compared to the contrastive pre-training, including the HWU64 dataset. Which shows that it is beneficial to include the target dataset during self-supervised contrastive learning. We leave whether self-supervised contrastive pre-training only on the target intent dataset benefits as a future study.

Is the training sensitive to hyper-parameters? We also study the effects of hyper-parameters of contrastive learning, *i.e.*, the temperature τ and weight λ'. We set $\tau \in \{0.05, 0.1, 0.3, 0.5\}$ and $\lambda' \in \{0.01, 0.03, 0.05, 0.1\}$. In our primary experiments, we do not find τ has a notable influence during the self-supervised contrastive pre-training on the first stage. Besides, we found that a batch size larger than 32 works well in the pre-training phase. However, during the few-shot fine-tuning stage, when setting τ to a small value 0.05, which heavily enforces the penalty to hard negative examples and λ' to a large value 0.1, which increases the weight of supervised contrastive learning loss, the performance drops significantly. In addition, the batch size influences performance on this stage. Therefore, few-shot supervised contrastive loss is sensitive to hyper-parameters when there are limited training examples. We leave more studies to future work.

7 Conclusion

In this paper, we improve the performance of few-shot intent detection via contrastive pre-training and fine-tuning. It first conducts self-supervised contrastive pre-training on collected intent detection datasets without using any labels, where the model implicitly learns to separate fine-grained intents. Then it performs the few-shot fine-tuning based on the joint intent classification loss and supervised contrastive learning loss, where the supervised contrastive loss encourages the model to distinguish intents explicitly. Experimental results on three challenging datasets show that our proposed method achieves state-of-the-art performance.

8 Acknowledgements

This work is supported in part by NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941. We thank the anonymous reviewers for their helpful and thoughtful comments.

References

Inigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. *ACL 2020*, page 38.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *EMNLP*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *EMNLP*, pages 2787–2792.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2. In *ACL*, pages 583–592.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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 *EMNLP*, pages 1311–1316.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021a. Fast, effective and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *arXiv preprint arXiv:2104.08027*.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*.

Zhiwei Liu, Yongjun Chen, Jia Li, Philip S Yu, Julian McAuley, and Caiming Xiong. 2021b. Contrastive self-supervised sequential recommendation with robust augmentation. *arXiv preprint arXiv:2108.06479*.

Zhiwei Liu, Ziwei Fan, Yu Wang, and Philip S Yu. 2021c. Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer. *arXiv preprint arXiv:2105.00522*.

Zhiwei Liu, Xiaohan Li, Ziwei Fan, Stephen Guo, Kannan Achan, and S Yu Philip. 2020. Basket recommendation with multi-intent translation graph neural network. In *IEEE Big Data*, pages 728–737. IEEE.

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

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020b. Example-driven intent prediction with observers. *arXiv preprint arXiv:2010.08684*.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained language models. *arXiv preprint arXiv:2004.13952*.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In *KDD*, pages 1150–1160.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *Image*, 2:T2.

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.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *SLT*, pages 19–24. IEEE.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020a. ToD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogues. *EMNLP*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020b. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Congying Xia, Caiming Xiong, and Philip Yu. 2021a. Pseudo siamese network for few-shot intent generation. In *ACM SIGIR*.

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

Congying Xia, Wenpeng Yin, Yihao Feng, and S Yu Philip. 2021b. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *NAACL-HLT*, pages 1351–1360.

Congying Xia, Chenwei Zhang, Hoang Nguyen, Jiawei Zhang, and Philip Yu. 2020b. Cg-bert: Conditional text generation with bert for generalized few-shot intent detection. *arXiv preprint arXiv:2004.01881*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *ACL*, pages 87–94.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *NeurIPS*, 33:5812–5823.

Jian-Guo Zhang, Kazuma Hashimoto, Yao Wan, Ye Liu, Caiming Xiong, and Philip S Yu. 2021. Are pretrained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. *arXiv preprint arXiv:2106.04564*.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, S Yu Philip, Richard Socher, and Caiming Xiong. 2020a. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *EMNLP*, pages 5064–5082.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, S Yu Philip, Richard Socher, and Caiming Xiong. 2020b. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *SEM, pages 154–167. Association for Computational Linguistics.