All Labels Together: Low-shot Intent Detection with an Efficient Label Semantic Encoding Paradigm

Jiangshu Du
University of Illinois at Chicago
jdu25@uic.edu

Congying Xia
Salesforce Research
c.xia@salesforce.com

Wenpeng Yin
Penn State University
wenpeng@psu.edu

Tingting Liang
Hangzhou Dianzi University
liangtt@hdu.edu.cn

Philip Yu
University of Illinois at Chicago
psyu@uic.edu

Abstract

In intent detection tasks, leveraging meaningful semantic information from intent labels can be particularly beneficial for few-shot scenarios. However, existing few-shot intent detection methods either ignore the intent labels, (e.g. treating intents as indices) or do not fully utilize this information (e.g. only using part of the intent labels). In this work, we present an end-to-end One-to-All system that enables the comparison of an input utterance with all label candidates. The system can then fully utilize label semantics in this way. Experiments on three few-shot intent detection tasks demonstrate that One-to-All is especially effective when the training resource is extremely scarce, achieving state-of-the-art performance in 1-, 3- and 5-shot settings. Moreover, we present a novel pretraining strategy for our model that utilizes indirect supervision from paraphrasing, enabling zero-shot cross-domain generalization on intent detection tasks. Our code is at https://github.com/jiangshdd/AllLablesTogether.

1 Introduction

Few-shot intent detection aims to identify the intents of user utterances with only a few labeled examples. Recent works can be mainly summarized into three categories: 1) **Standard classifier-based approaches** (He et al., 2022a,b; Zhang et al., 2022, 2021), which leverage pretrained language models (PLMs) equipped with a standard classifier layer (e.g., MLP), treating intent labels as indices; 2) **Example-based approaches** (Zhang et al., 2020; Mehri and Eric, 2021; Vulić et al., 2021), which learn to compare the similarities between different examples and classify an input utterance based on the closest neighbor in the training data; 3) **Intent semantic aware approaches** (Qu et al., 2021; Xia et al., 2021; Du et al., 2022), which explicitly incorporate intent label words during training. However, both classifier-based and example-based methods disregard label semantic information, which is an important source of supervision in few-shot scenarios. Exiting intent semantic aware approaches also suffer from different drawbacks, such as relying on large-scale pretraining datasets and only partially using intent labels. More details on related work are discussed in Appendix A.

To solve these issues, we propose an end-to-end intent semantic aware model, One-to-All. It concatenates each utterance with the entire intent label set as the input and then encodes them simultaneously. In this way, the semantic information of all intents is fully utilized and integrated with utterances. The encoded embeddings of labels and utterances are subsequently used for contrastive learning. We define a new contrastive learning paradigm by comparing the representations of utterances and intents directly. This approach ensures utterances are moved closer to their gold intents while distancing them from any incorrect ones. Furthermore, we introduce a novel pretraining strategy for One-to-All that leverages indirect supervision from paraphrase identification datasets. Through this strategy, the model develops the ability to understand semantic similarities and distinctions among sentences, generalizing its comprehension to unseen intents in zero-shot intent detection tasks.

To demonstrate the effectiveness of our proposed model, we conduct experiments on three fine-grained intent detection tasks: BANKING77 (Casanueva et al., 2020), HWU64 (Liu et al., 2019a) and CLINC150 (Larson et al., 2019), under low-shot settings (0-, 1-, 3- and 5-shot). The results show that One-to-All is especially effective in extreme few-shot scenarios, with an average improvement of 4.62% in 1-shot and 2.60% in 3-shot settings over the state-of-the-art (SOTA) without any pretraining. Our model also achieves SOTA in 5-shot scenarios with pretraining on out-of-domain (OOD) data. Furthermore, One-to-All shows great cross-domain generalization capabilities in the zero-shot setting when further pretrained on

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 131–138
November 1–4, 2023. ©2023 Association for Computational Linguistics

paraphrasing identification datasets.

Our contributions can be summarized as follows. First, we proposed an end-to-end `One-to-All` system that enables the comparison of an input utterance with all label candidates via a newly defined contrastive learning paradigm. To our knowledge, this is the first work that can encode the entire label space while modeling the intent identification problem. Second, we go beyond few-shot intent detection and further achieve zero-shot cross-domain generalization with a novel pretraining stage of our model: pretraining on paraphrase identification datasets. This is the first work that effectively uses indirect supervision from paraphrasing to handle zero-shot intent identification tasks.

2 Methods

2.1 `One-to-All` Input Sequence Construction

Incorporating additional labels as contexts alongside utterances within an input sequence can help the model make a better decision (Du et al., 2022). `One-to-All` concatenates each utterance with the complete intent label set as the input and encodes them together. However, the limitation of the maximum input sequence length makes it impractical to include all labels in a single sequence, especially when the label space is large. Therefore, given an intent detection task with n intents, we take every k intents as a group and then all the intents will be divided into $m = \lceil n/k \rceil$ groups. To keep each group with a consistent number of elements, we introduce a special placeholder token <*plh*>. For the group whose number of intents (s) is less than k, we fill it with $(k - s)$ <*plh*>s to maintain consistency. Each utterance U is then duplicated m times and appended with the m groups of intents, respectively, yielding m input sequences, as shown in Figure 1(A). In this way, the model works on the entire intent space and incorporates multiple intents into a single sequence. In practice, we choose k by minimizing $n \bmod k$ to avoid introducing too many <*plh*>s. Each input sequence is then shuffled k times during training to perform data augmentation, which is beneficial for few-shot tasks.

2.2 Intent-Aware Contrastive Learning

To better exploit intent semantic information, we apply contrastive learning (CL) between utterances and intents. Previous works perform CL between input texts and their augmented views (Gao et al., 2021; Yan et al., 2021; Chen et al., 2020; Mou et al., 2022), or instances under different classes (Zhang et al., 2021; Gunel et al., 2021; Lin et al., 2022), while we directly perform CL between input texts (utterances) and classes (intents) since intents usually contain useful semantic meanings. Therefore, the representations of intents can be regarded as different views of utterances or explicit cluster centroids. As shown in Figure 1 (B), we feed the input sequence constructed from Section 2.1 into RoBERTa and obtain a list of token-level representations. Then for the utterance U and all the intents $I_j, j \in \{1, ..., k\}$ in the input sequence, we average the representations of their corresponding tokens as their representations z_U and $z_j, j \in \{1, ..., k\}$. Finally, we use a shared MLP projector to project z_U and z_j into the same semantic space, yielding the final utterance representation h_U and intent representation h_j. We use h_U^+ to denote the representation of the gold intent I_U for U. Let $sim(u, v)$ denote the cosine similarity between u and v. The contrastive learning loss is then defined as follows:

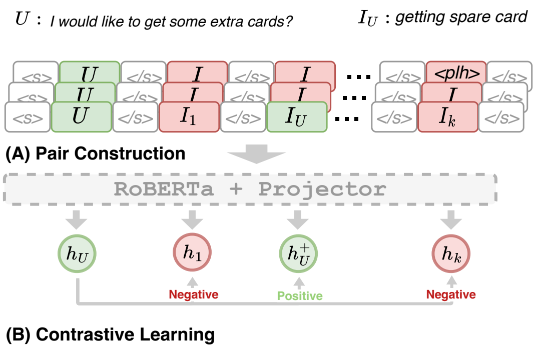


Figure 1: Overview of `One-to-All`. (A) shows the input sequence construction strategy. I_U indicates the gold intent of U. (B) shows positive and negative instances in contrastive learning.

$$l = -\frac{1}{N_b} \sum_{i=1}^{N_b} log \frac{e^{sim(\mathbf{h}_U^i, \mathbf{h}_U^{i+})/\tau}}{\sum_{j=1}^{k} e^{sim(\mathbf{h}_U^i, \mathbf{h}_j^i)/\tau}} \quad (1)$$

, where N_b is the batch size. h_*^i means the representation in the i-th input in a batch. τ is the temperature parameter. Through the contrastive loss, `One-to-All` pushes utterances closer to their gold intents and meanwhile drives them away from all the incorrect intents. For the input sequences that do not include gold intents, we simply set the numerator of Equation 1 to 1 so the model only pushes the utterance away from other intents.

By shuffling the order of concatenated intents within the sequence, a large number of training instances and distinct contrastive pairs can be generated. This is particularly beneficial for both few-shot and contrastive learning. Furthermore, a more robust model is yielded since we enforce the model to recognize the correct intent regardless of its location in the concatenated input sequence.

2.3 Zero-shot Intent Detection via Paraphrase Identification Pretraining

Paraphrase identification (Socher et al., 2011) is a task that aims at identifying if two sentences have the same meaning. Pretraining on the paraphrase detection identification dataset encourages models to capture the essence of utterances rather than relying solely on surface-level patterns. This can improve the ability of `One-to-All` to distinguish between similar intents that might have subtle differences in wording or phrasing. Therefore, we propose a novel pretraining stage to effectively use indirect supervision from paraphrasing for zero-shot intent detection task. Specifically, for each sentence, we regard its paraphrase as its gold label and select other t most similar sentences with the help of Sentence-BERT (Reimers and Gurevych, 2019) as the negative labels. For the target task with n intents, we set $t = n - 1$ so each sentence has totally n labels to compare as well. We set k, which is the number of intents in each group, the same as the target task to keep the pretraining and downstream tasks consistent.

Pretraining on out-of-domain (OOD) data can further improve the performance of `One-to-All`. Given an intent detection task, we treat other data which do not have domain overlapping with this task as the OOD data. During OOD pretraining, the intent space is a union of intents from all the OOD data. We follow the same sequence construction strategy and training process stated in Section 2.1 and Section 2.2.

3 Experiments

We conduct experiments on three intent detection tasks under both few-shot and zero-shot settings.

Datasets. Our model is evaluated on three fine-grained intent detection datasets: BANKING77 (Casanueva et al., 2020), HWU64 (Liu et al., 2019a), and CLINC150 (Larson et al., 2019). Each dataset contains one or multiple domains and a fine-grained intent space. Dataset statistics are shown in Appendix Table 2. We randomly sample 10% training data as the *dev set*, following Zhang et al. (2021) and Mehri et al. (2020). For each target task, we use the other two datasets excluding the similar domains and intents as its OOD pretraining data. For example, we take BANKING77 as the target task, which contains banking-related intents. To obtain its OOD pretraining data, we combine the data from HWU64 and CLINC150 but remove the "Banking" and "Credit Cards" domains. The statistics of the OOD data for each target task are shown in Appendix Table 3. We also sample 10% OOD data as the *dev set* for the OOD pretraining.

For paraphrase detection pretraining, we leverage the public Quora Question Pairs (QQP) dataset[1]. To incorporate more sentences in a single pair, we filter short sentence pairs from QQP by setting the max number of words and characters in a sentence as 10 and 40, respectively. The filtered QQP dataset contains 31,412 paraphrase pairs.

Baselines. We compare our method with six baselines in the three categories as we described in Section 1. Standard classifier-based baselines: 1) RoBERTa with a standard classifier. 2) CPFT (Zhang et al., 2021) performs self-supervised CL on multiple intent detection datasets and applies supervised CL on the target task. 3) RoBERTa-SPI (Zhang et al., 2022) introduces two regularizers to improve supervised pretraining via isotropization. For example-based approaches, we consider its prior SOTA model: 4) DNNC (Zhang et al., 2020), which identifies intents by finding the nearest neighbors of utterances in the training set and is pretrained on three natural language inference datasets. For intent semantic aware methods, we compare: 5) Context-TE and 6) Parallel-TE (Du et al., 2022). These approaches incorporate multiple intents into a single textual entailment sequence. Context-TE relies on indirect supervision from MNLI (Williams et al., 2018). Parallel-TE is more comparable to `One-to-All`, it also encodes utterances and intents simultaneously, but it is a pipeline model that only selects top-k intents for each utterance. Among all the baselines, CPFT and DNNC are the prior SOTA models for 5- and 10-shot settings. Our paper considers more challenging scenarios, specifically the 1- and 3-shot settings. The implementation of the baselines and our model are detailed in Appendix C.

Results. For few-shot experiments, we conduct three runs with distinct training data samples, following Du et al. (2022). Table 1 shows the average

[1] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

Model	BANKING77			HWU64			CLINC150		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
RoBERTa	35.31(2.22)	64.78(0.76)	75.47(1.69)	40.34(2.66)	67.44(1.37)	75.71(1.26)	48.48(1.03)	80.91(2.10)	86.82(1.40)
CPFT	40.71(2.25)	71.57(0.34)	79.73(0.52)	50.61(2.18)	72.91(2.32)	79.82(1.64)	58.58(0.57)	83.12(0.43)	90.60(0.70)
RoBERTa-SPI	43.81(2.01)	65.63(0.66)	72.20(1.45)	53.75(2.13)	70.95(1.14)	75.58(1.24)	67.57(0.99)	83.31(1.78)	87.76(1.08)
DNNC	30.23(2.51)	72.21(1.02)	79.94(1.77)	29.77(1.45)	75.25(2.69)	79.31(0.19)	30.17(2.33)	87.07(0.44)	90.44(1.03)
Context-TE	64.22(0.50)	73.27(0.43)	77.07(0.57)	64.39(1.64)	73.45(1.72)	78.16(0.94)	74.71(0.91)	84.56(1.54)	87.61(0.79)
Parallel-TE	64.34(1.23)	72.20(1.00)	76.23(0.37)	61.96(0.46)	74.10(0.51)	78.10(1.40)	74.54(0.81)	83.66(0.84)	86.61(0.58)
`One-to-All`	**66.36**(0.46)	**76.13**(0.45)	79.75(0.78)	**68.77**(1.94)	**78.16**(2.04)	**79.89**(0.30)	**77.63**(0.63)	**87.09**(1.44)	89.88(0.81)
w/ OOD	**67.93**(0.28)	**76.92**(0.18)	**80.51**(0.88)	**73.17**(0.37)	**79.95**(0.68)	**82.50**(1.05)	**79.21**(0.43)	**88.01**(1.46)	**90.76**(0.63)

Table 1: Test accuracy (%) and standard deviation on three dataset under three few-shot scenarios. The first and second highest results are formatted in **bold** and underline, respectively.

accuracy and standard deviation on three datasets under 1-, 3-, and 5-shot settings. `One-to-All` outperforms all the baselines remarkably in 1- and 3-shot scenarios across three datasets without any pretraining. For example, `One-to-All` improves the state-of-the-art result for 1-shot on HWU64 by 6.81%. After pretraining on OOD data, the improvement percentage increases to 13.64%. For the 5-shot setting, `One-to-All` achieves comparable results without pretraining and outperforms all the baselines with pretraining on OOD. The example-based model DNNC performs extremely poorly on 1-shot tasks even though it achieves good performances in 5-shot, showing its limitation when training resources are extremely scarce.

For the zero-shot setting, we first did preliminary experiments evaluating all baselines and our model on the target tasks without extra pretraining. The results are reported in Appendix Table 4. Despite poor performance across all models, `One-to-All` remains the top-performing approach. Then we pretrain the model on OOD/QQP data without accessing any in-domain training data. The results are shown in Figure 2. The zero-shot performance of `One-to-All` exhibits significant improvement following pretraining on QQP data, indicating the effectiveness of our novel pretraining strategy. Comparing the zero-shot results with the few-shot results in Table 1, we can observe that the performance of `One-to-All` (OOD) under the zero-shot setting even outperforms some baselines (RoBERTa, CPFT, and DNNC) under the 1-shot setting. Thus, we compare our model, `One-to-All`, with the two strongest baselines, Context-TE and Parallel-TE, in the zero-shot setting. As shown in Figure 2, `One-to-All` (OOD) outperforms both Context-TE (OOD) and Parallel-TE (OOD) in most settings. The zero-shot performance of `One-to-All` is further boosted by pretraining on both OOD and QQP data, outperforming both Context-TE and Parallel-TE by a large margin. Once again, this finding highlights the effectiveness of utilizing indirect supervision from paraphrase identification datasets.

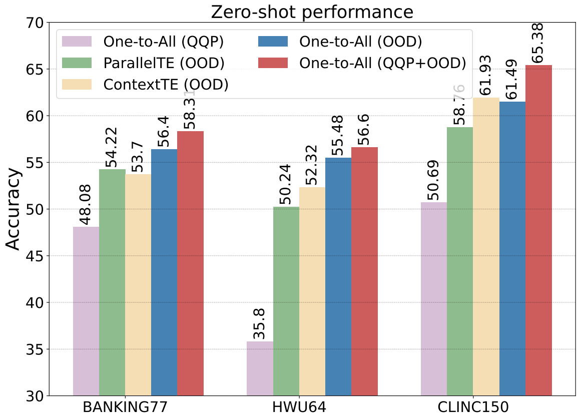


Figure 2: Zero-shot performance. OOD or QQP indicates that the model is pretrained on OOD or QQP data.

Analysis We investigate how our end-to-end design impacts our model performance by comparing it with Parallel-TE through a case study on BANKING77. We run `One-to-All` and Parallel-TE under the 3-shot setting three times. The top-k filtering step in Parallel-TE misses gold intents for 89 utterances in the *test set*, while `One-to-All` can correctly predict 34.7 of them on average, which brings a 1.1% improvement given the size of *test set* is 3080. This suggests that our end-to-end design effectively identifies partial intents that may have been overlooked by the pipeline system.

4 Conclusion

In this paper, we propose an end-to-end intent semantic aware intent detection model `One-to-All` to fully leverage intent semantics via contrastive learning. Experiments show that it is especially effective when training resource is scarce.

Limitations

Although we perform OOD pretraining on our model and gain performance improvement, the OOD data we use in our experiment is only from at most two datasets. There are many publicly available intent detection datasets from different domains, such as ATIS (Hemphill et al., 1990) and SNIPS (Coucke et al., 2018), can be used for the OOD pretraining. We believe pretraining on large-scale OOD datasets can further boost the performance of our model, and we leave it for our future work.

Acknowledgments

The authors appreciate the reviewers for their insightful comments and suggestions. This work is supported in part by NSF under grant III-2106758.

References

Pavel Burnyshev, Andrey Bout, Valentin Malykh, and Irina Piontkovskaya. 2021. InFoBERT: Zero-shot approach to natural language understanding using contextualized word embedding. In *Proceedings of RANLP*, pages 208–215.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, volume 119, pages 1597–1607.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190.

Jiangshu Du, Wenpeng Yin, Congying Xia, and Philip S. Yu. 2022. Learning to select from multiple options. *ArXiv*, abs/2212.00301.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proceedings of ICLR*.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. SPACE-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. In *Proceedings of COLING*, pages 553–569.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of SIGIR*, page 187–200.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990.*

Dmitry Lamanov, Pavel Burnyshev, Katya Artemova, Valentin Malykh, Andrey Bout, and Irina Piontkovskaya. 2022. Template-based approach to zero-shot intent recognition. In *Proceedings of INLG*, pages 15–28.

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 EMNLP-IJCNLP*, pages 1311–1316.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of NAACL*, pages 2543–2556.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. *ArXiv*, abs/1903.05566.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of NAACL-HLT*, pages 2979–2992.

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

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022. Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning. In *Proceedings of ACL (Short Papers)*, pages 46–53.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. 2022. Label semantic aware pre-training for few-shot text classification. In *Proceedings of ACL*, pages 8318–8334.

Jin Qu, Kazuma Hashimoto, Wenhao Liu, Caiming Xiong, and Yingbo Zhou. 2021. Few-shot intent classification by gauging entailment relationship between utterance and semantic label. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Richard Socher, Eric Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NeurIPs*, pages 801–809.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of EMNLP*, pages 1151–1168. 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 NAACL-HLT*, pages 1112–1122.

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 EMNLP*, pages 917–929.

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

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of ACL*, pages 5065–5075.

Haode Zhang, Haowen Liang, Yuwei Zhang, Li-Ming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Lam. 2022. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *Proceedings of NAACL-HLT*, pages 532–542.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of EMNLP*, pages 1906–1912.

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 EMNLP*, pages 5064–5082.

A Related Work

Recent research on low-shot intent detection can be broadly classified into three main categories: standard classifier-based, example-based, and intent semantic aware approaches.

Standard-classifier methods discard intent semantics and usually require an extra pretraining process on the extra corpus. For example, He et al. (2022a,b); Wu et al. (2020) pretrain models on large-scale dialog datasets. Zhang et al. (2021) pretrain a RoBERTa model on various intent datasets via self-supervised contrastive learning. Zhang et al. (2022) propose two regularizers to improve its supervised pretraining step via isotropization.

Example-based approaches aim to learn the similarities between various examples and classify an input utterance by identifying its closest neighbor in the training data. For instance, Zhang et al. (2020); Mehri and Eric (2021) determines the intent of an utterance by searching for its nearest neighbors among all training utterances. These approaches also ignore the semantic information of intents.

Existing intent semantic aware approaches also exhibit various drawbacks and limitations. For example, LSAP (Mueller et al., 2022) incorporates intent semantics into generative models via pretraining. More specifically, during the pretraining stage, LSAP takes partially masked utterance-intent pairs as the input and predicts the masked contents. However, this approach relies on large-scale pretraining data to obtain decent performance. Qu et al. (2021) and Xia et al. (2021) cast ID as textual entailment (TE), treating utterances and intents as premises and hypotheses, respectively. But these two models are only able to compare one single utterance with one single intent, which makes them unaware of other intent options. Du et al. (2022) learn to select the best intent for an utterance by providing the top-k intents for that utterance in one training example. Despite providing a one-to-many comparison, Du et al. (2022) are only able to view the top-k intents rather than the entire intent label set during training. Additionally, Du et al. (2022) propose a pipeline model. Their performance is constrained by the accuracy of the first stage of the pipeline, which is identifying the top-k intents. Moreover, Lamanov et al. (2022) propose a template-based approach for modeling intents and utterances as sentence pairs. Burnyshev et al. (2021) use a deep contextualized model to embed utterances and the natural language descriptions of user intents in zero-shot scenarios.

Different from all the literature we discussed above, we propose an end-to-end intent semantic aware system. By fully utilizing label semantics via contrastive learning, our model achieves SOTA performance even without pretraining on additional datasets.

B Dataset Statistics

Dataset	Domain	Utterance	Intent
BANKING77	1	13,083	77
HWU64	21	10,030	64
CLINC150	10	22,500	150

Table 2: Dataset statistics.

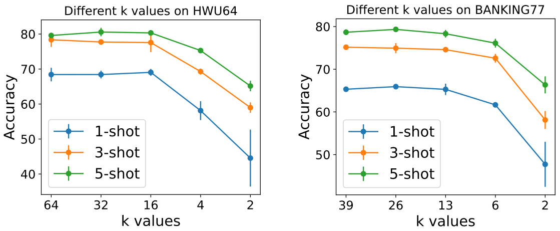
Target	Domain	Utterance	Intent
BANKING77	29	28,030	183
HWU64	11	35,453	225
CLINC150	21	9,854	63

Table 3: Statistics of the OOD pretraining data used for each target task.

C Implementation Details.

All the baselines and `One-to-All` adopt RoBERTa-base (Liu et al., 2019b) as the backbones for a fair comparison.

Baseline implementation. For the RoBERTa model, we implement it with the Hugging Face[2] library. For RoBERTa-SPI, DNNC, Context-TE, and Parallel-TE, we directly run their open-source code under our experiment settings. It is important to note that the original Context-TE and Parallel-TE models utilize RoBERTa-large as their base models. However, in our implementation, we replace it with RoBERTa-base to ensure a fair comparison. Regarding CPFT, we strive to replicate its methodology to the best of our ability, given the unavailability of its code and pretrained checkpoints to the public. The original paper trained CPFT on a fixed set of 5-shot data, while in our experiments, we conduct three runs with uniquely sampled few-shot training data, which mitigates the potential influence of data sampling bias.



(a) HWU64 (b) BANKING77

Figure 3: Model performance with different k values on the *dev set* of HWU64 and BANKING77.

Model	BANKING	HWU	CLINC
RoBERTa	1.65	1.77	2.01
CPFT	1.91	2.03	2.78
RoBERTa-SPI	1.90	1.87	2.52
Context-TE	1.04	1.20	1.43
Parallel-TE	1.66	2.14	2.24
`One-to-All`	**6.04**	**13.05**	**5.84**

Table 4: Zero-shot performance (%) without pretraining on additional datasets. DNNC is not included since it requires at least one training example.

`One-to-All` implementation. For few-shot tasks and OOD pretraining, we train the model for 10 and 3 epochs, respectively, and keep the best ones on the *dev set*. For paraphrase detection pretraining, we train the model for 3 epochs. All the training batch size is set to 8 and the learning rate is 2e-5. The output dimension of the MLP projector is set to 768 and the temperature parameter $\tau = 0.1$. We set k to 26, 32, 30 for BANKING77, HWU64, and CLINC150, respectively, according to the observations on their *dev sets*. Details are discussed in Appendix E.

D Zero-shot Preliminary Experiments

E Additional Analysis

We try to explore how the number of intents in each input sequence, k, influences the performance. We explore the influence of k by conducting experiments on the *dev set* of HWU64 and BANKING77 with different k values. As shown in Figure 3, the model performance stays similar when $k > 20$ but drops sharply when k is below 10. This is probably due to the reduction of contrastive instances in a

[2] https://huggingface.co/

single pair. Therefore, we set k to 26, 32, 30 for BANKING77, HWU64, and CLINC150, respectively, and it can bring two benefits: 1) it is more friendly for the paraphrase detection pretraining as it is the max number of the sentences that can be incorporated into a single pair; 2) it minimizes $n \bmod k$ as we discussed in Section 2.1.