2023 IEEE 17th International Conference on Semantic Computing (ICSC) | 978-1-6654-8263-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSC56153.2023.00014

Robust Zero-Shot Intent Detection via Contrastive Transfer Learning

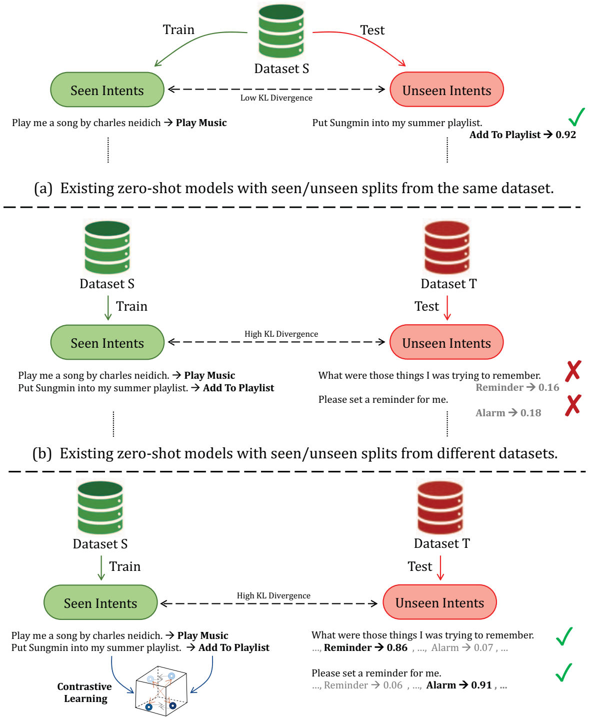
M.H. Maqbool♣, F.A. Khan♣, A.B. Siddique♡, Hassan Foroosh♣
hasanmaqbool@knights.ucf.edu♣, Fahad.Khan@ucf.edu♣, siddique@cs.uky.edu♡, Hassan.Foroosh@ucf.edu♣

University of Central Florida♣, University of Kentucky♡

Abstract*—Intent detector is a central component of any task-oriented conversational system. The goal of the intent detector is to identify the user's goal by classifying natural language utterances. In recent years, research has focused on supervised intent detection models. Supervised learning approaches cannot accommodate *unseen intents*, which may emerge after the system has been deployed — the more practically relevant setting, known as zero-shot intent detection. The existing zero-shot learning approaches split a dataset into seen and unseen intents for training and evaluations without taking the sensitivity of the data collection process into account. That is, humans tend to use repeated vocabulary and compose sentences with similar compositional structures. We argue that the source-to-target relationship learning objective of zero-shot approaches under typical data split procedure renders the zero-shot models prone to misclassifications when target intents are divergent from source intents. To this end, we propose INTEND, a zero-shot INTENt Detection methodology that leverages contrastive transfer learning and employs a zero-shot learning paradigm in its true sense. First, in contrast to partitioning the training and testing sets from the same dataset, we demonstrate that selecting training and testing sets from two different datasets, allows for rigorous zero-shot intent detection evaluations. Second, our employed contrastive learning goal encourages the system to focus on learning a generic similarity function, rather than on commonly encountered patterns in the training set. We conduct extensive experimental evaluations using four public intent detection datasets for up to 150 unseen classes. Our experimental results show that INTEND consistently outperforms state-of-the-art zero-shot techniques by a substantial margin. Furthermore, our approach achieves significantly better performance than few-shot intent detection models.

I. Introduction

Identifying users' intent from natural language utterances is a crucial step for conversational systems. For example, a conversational system can be issued a command, *"Set up a reminder for grocery on April, 30 at 3pm"*. Accurately recognizing the user's intent (i.e., *"SetUpReminder"* in this example) enables the system to execute the necessary steps to set the reminder. In a supervised setting [1], [2], [3], a model can be trained on labeled training data, leading to high performance on the intent detection task. However, the challenge arises when new unseen intents emerge in the operational lifespan of a conversational system. Emerging intents necessitate their flexible accommodation without sacrificing the system's performance on the critical task. The supervised learning models fall short of providing the robust absorption of nascent intents. Since supervised models tend to learn a probability distribution over intent labels, they are unable to integrate new intents without (re-)training the model on the (new) labeled dataset. Moreover, acquiring labeled training data for each new intent is laborious and expensive which motivates the zero-shot intent detection task [4].


Train
Test
Dataset S
Seen Intents
Unseen Intents
Low KL Divergence
Play me a song by charles neidich → Play Music
Put Sungmin into my summer playlist.
Add To Playlist → 0.92
✓

(a) Existing zero-shot models with seen/unseen splits from the same dataset.

Dataset S
Dataset T
Train
Test
Seen Intents
Unseen Intents
High KL Divergence
Play me a song by charles neidich. → Play Music
Put Sungmin into my summer playlist. → Add To Playlist
What were those things I was trying to remember.
Reminder → 0.16
✗
Please set a reminder for me.
Alarm → 0.18
✗

(b) Existing zero-shot models with seen/unseen splits from different datasets.

Dataset S
Dataset T
Train
Test
Seen Intents
Unseen Intents
High KL Divergence
Play me a song by charles neidich. → Play Music
Put Sungmin into my summer playlist. → Add To Playlist
Contrastive Learning
What were those things I was trying to remember.
..., Reminder → 0.86 , ..., Alarm → 0.07 , ...
✓
Please set a reminder for me.
..., Reminder → 0.06 , ..., Alarm → 0.91 , ...
✓

(c) Proposed zero-shot approach with seen/unseen splits from different datasets.

Fig. 1: (a) The typical train and test splits for zero-shot intent detection models. (b) Existing zero-shot intent detection models show poor performance when confronted with divergent unseen intent labels at inference time. (c) The proposed, contrastive learning-based approach, provides the model with robust zero-shot adaptations for unseen intents.

Recently, several zero-shot (and few-shot) learning approaches have been devised with promising results [5], [6], [7]. A common zero-shot setting designates a fraction of classes in a dataset as unseen, while instances from the remaining classes are used for training. A model trained in this setting captures distribution-specific features from seen intents and,

978-1-6654-8263-9/23/$31.00 ©2023 IEEE
DOI 10.1109/ICSC56153.2023.00014

generally, performs well on unseen intents. Figure 1 (a) presents a scenario, where a dataset is split into seen and unseen intent classes. The seen intent class label “Play Music” has high similarity with an unseen class label “Add To Playlist”, which enables the zero-shot model to perform well on the unseen classes after being trained using the examples of seen classes. However, such models are not robust when encountered with divergent unseen classes from the seen classes. For example, Figure 1 (b) shows a scenario where seen and unseen classes are drawn from two different datasets. Naturally, intent classes along with the domain, vocabulary, and sentence structure, among other features, in both datasets, have a high probability to be diverse, posing a challenge for zero-shot models in this *true* zero-shot setting. The seen classes, “Play Music” and “Add To Playlist” have low similarity with unseen classes, “Reminders” and “Alarms”, and existing zero-shot models are shown to demonstrate poor performance in this setting. Intuitively, if the dataset is homogeneous and exhibits low KL divergence between seen and unseen class distributions (e.g., Figure 1 (a)), the model could use the learned features for accurate inference. However, such guarantees are unlikely to exist in the real world where data exhibits high KL divergence and new intents may appear with a weaker affinity for already learned features (i.e., less similarity with seen utterances). We argue that the misclassifications in zero-shot models (e.g., Figure 1 (b)) are caused by the learning objective of these techniques – the source-to-target relationships.

In this work, we propose a robust zero-shot setting to address the misclassification issue inherent in typical zero-shot settings. First, we propose that the train and test datasets should be fully distinct, having different distributions, with a disjoint (or even overlapping) set of intents (as shown in Figure 1 (b) and (c)). This is a more challenging, practically relevant, and close-to real-world setting. To the best of our knowledge, this is the *first* work to propose conducting zero-shot intent detection experiments in this setting. Second, instead of a source-to-target relationship-based objective, a contrastive learning objective is adapted. The contrastive learning is self-supervised as well as task-independent, thus forcing the model to learn a generic embedding space (i.e., irrespective of seen or unseen class labels) where similar utterance and intent pairs are close to each other and dissimilar ones are far from one another. Third, language models trained on large-scale datasets, such as Sentence-BERT [8], are employed to capture contextual correspondences between utterances and intents. These models have millions of parameters and thus have the capability to generate rich representations even for new unseen class labels. Since datasets (train and test) are supposed to have heterogeneous distributions, the proposed zero-shot intent detection model minimizes the model’s predisposition to converge to observed classes, directing the learning process more towards generalization.

Figure 1 presents an overview of the proposed zero-shot intent detection approach. Assume our train dataset is S and our test dataset is T. The goal is to transfer generic features from S to T. It is to be noted that S and T do not necessarily have to have a disjoint set of intents The model is trained on S in a contrastive fashion, with N negative examples per positive instance, and validated on T with contrastive sampling. This training strategy imposes convergence criteria on the model’s training with respect to test dataset T that accommodates new intents in the dataset T. Because of the contrastive training objective, our model integrates newly emerging intents in a flexible manner without jeopardizing the system’s robustness. Specifically, using the dataset S, our training goal is to learn positive and negative associations between utterances and intents. It is important to mention that gradient updates are never performed on the dataset T. We show empirically that our learning objective instructs the model to learn general patterns that are then applied to the target dataset T.

We use four public intent detection benchmarks to demonstrate the effectiveness of our proposed approach. We conduct experimental evaluations in the (true) zero-shot settings and the results show that our proposed zero-shot intent detection model, INTEND, outperforms state-of-the-art models in a wide range of experiments. Furthermore, we conduct experiments using a few-shot setting and results demonstrate that our model’s quantitative performance is better or competitive with several state-of-the-art large models, such as DNNC [9] and DialoGLUE (ConvBERT-DG) [10].

Our contributions are summarized as follows:

- We effectively couple pre-trained language models with contrastive transfer learning for zero-shot intent detection.
- We show that INTEND accommodates newly emerging intents with high F1 score and outperforms state-of-the-art models in zero-shot and few-shot intent detection settings.
- We also present a distribution divergence analysis between training and testing datasets to demonstrate a rigorous zero-shot evaluation setting.

II. Preliminaries

A. Problem Formulation

Suppose we have the training and testing datasets denoted by $S = (X^s, I^s)$ and $T = (X^t, I^t)$, respectively, where X^s and X^t denote the set of training and test utterances, respectively. Similarly, I^s and I^t represent the set of training and test intents, respectively. We further define $X^s = \{\mathcal{X}_1^s, \mathcal{X}_2^s, \ldots, \mathcal{X}_n^s\}$ and $I^s = \left\{\mathcal{I}_1^s, \mathcal{I}_2^s, \ldots, \mathcal{I}_p^s\right\}$ where n is the number of utterances and p denotes the number of intents in training dataset S. Similarly, utterance/intent pairs for test dataset T can be defined as $X^t = \{\mathcal{X}_1^t, \mathcal{X}_2^t, \ldots, \mathcal{X}_m^t\}$ and $I^t = \left\{\mathcal{I}_1^t, \mathcal{I}_2^t, \ldots, \mathcal{I}_q^t\right\}$ where q and m denote the number of intents and utterances, respectively. $\bar{\mathcal{N}}_i^s$ represents the set of negative samples for a positive sample $\mathcal{X}_i^s$ consisting of negative utterance/intent pairs of the form $(\mathcal{X}_i^s, \bar{\mathcal{I}}_K^s)$ where $\bar{\mathcal{I}}_K^s$ represent an intent not related to $\mathcal{X}_i^s$ and K is the number of negative samples. Having a model trained on training dataset S, our objective is to classify a test utterance $\mathcal{X}_i^t$ belonging to the test dataset T. We emphasize that $I^s \cap I^t = \phi$ is not a requirement for our approach to work. This setting is referred to as generalized zero-shot intent detection.

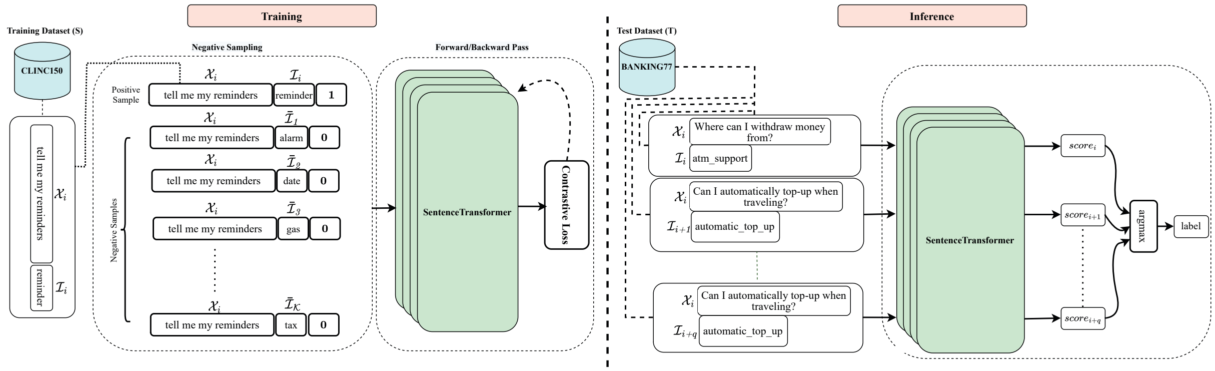


Fig. 2: The training and inference procedures along with the end-to-end architecture of our proposed framework.

B. *Unsupervised Representation Learning*

The unsupervised representation learning approaches have revolutionalized the field of natural language processing (NLP). Given the abundance of textual data, the unsupervised learning techniques [11], [12], [13], [14], [15], [16] allowed for the development of pre-trained language models. In this work, we use pre-trained language models and fine-tune them for capturing generic patterns that can effectively be transferred to the test dataset. We learn sentence level representations using a variation of Sentence-BERT [8] that uses MPNet [15] instead of BERT [13]. MPNet [15] builds upon the strengths of BERT [13] and XLNet [14], addresses their issues, and outperforms them on various NLP tasks. In the following, we provide a brief overview of both modeling approaches.

Masked Language Model (MLM). BERT [13] employs transformer architecture [17] and MLM for learning bidirectional representations. Consider $x = \{x_1, x_2, x_3, \ldots, x_n\}$ is a given sequence, where 15% tokens are masked (i.e., replaced with $[MASK]$ token). Let us represent the masked tokens with $\mathcal{M}$ and set of masked tokens by $x_{\mathcal{M}}$, where $x_{\setminus\mathcal{M}}$ is the sequence after masking. MLM objective can be written as:

$$\log P\left(x_{\mathcal{M}}|x_{\setminus\mathcal{M}};\theta\right) \approx \sum_{m\in\mathcal{M}} \log P(x_m|s_{\setminus\mathcal{M}};\theta) \qquad (1)$$

Permuted Language Model (PML). XLNET [14] proposes PLM technique in order to utilize the power of bidirectional context and autoregressive modeling. Consider an input sequence $x = \{x_1, x_2, x_3, \ldots, x_n\}$. Let us suppose that the set Z_n represents all the $n!$ permutations of the input sequence. Let $z_n \in Z$ be one of the permutations and the current token to be predicted is located at index t, then $z_{<t}$ are the first $t-1$ tokens. As a concrete example, let us say $z_n = \{1, 3, 5, 2, 4\}$; if $t = 4$ then $z_t = 2$, $x_{z_t} = x_2$ and $z_{<t} = \{1, 3, 5\}$. Keeping this notation in view, PLM formulation is given below:

$$\log P(x;\theta) = \mathbb{E}_{z\in Z_n} \sum_{t=c+1}^{n} \log P(x_{z_t}|x_{z<t};\theta) \qquad (2)$$

where c represents non-predicted tokens. $x_{z_{<=t}}$

III. Our Approach

Our proposed framework, INTEND, employs a variant of Sentence-BERT [8] that uses pre-trained MPNet [15] as the backbone model and further fine-tunes it with a contrastive learning strategy to achieve zero-shot generalization. An End-to-end training and inference pipeline of INTEND is presented in Figure 2. Given a test utterance $\mathcal{X}_i^t$, our goal is to assign an intent from an unseen set of intents I^t to the utterance. It is important to recall that the model is trained only using examples from training dataset $S = (X^s, I^s)$. In the following, we first explain the crucial building blocks of INTEND, then provide details about contrastive learning.

A. *Building Blocks*

MPNet. Even though BERT is a powerful model, it suffers by not considering the masked token dependencies. On the other hand, XLNet captures bidirectional contexts by considering all permutations of factorization order, but it does not consider the full position information of a sentence and thus suffers from *pretrain-finetune* discrepancy. MPNet [15] leverages the best of MLM (i.e., BERT) and PLM (i.e., XLNet) modeling by proposing a unified view and by addressing their limitations. For a given sequence $x = \{x_1, x_2, x_3, x_3, x_4, x_5\}$ where tokens x_2 and x_4 are masked. This sequence is permuted to get $x = \{x_1, x_3, x_5, x_2, x_4\}$ with the masked tokens at the end of the sequence. Since transformer architecture is independent of input order as long as tokens and their positions are correctly associated [15], the objective of MLM can be rewritten as:

$$\mathbb{E}_{z\in Z_n} \sum_{t=c+1}^{n} logP(x_{z_t}|x_{z<=c}, M_{z_{>c}}, ;\theta) \qquad (3)$$

In our given permuted sequence $n = 5$, $c = 3$, to be predicted tokens are $x_{z>c}$ and $M_{z_{>c}}$ are the masked tokens. We can notice that MLM's unified view formulation and PLM's formulation from Equation 2 are similar. Finally, MPNet's formulation can be given as follows:

$$\mathbb{E}_{z\in Z_n} \sum_{t=c+1}^{n} logP(x_{z_t}|x_{z<t}, M_{z_{>c}}, ;\theta) \qquad (4)$$

The pre-trained MPNet is fine-tuned on over 1 billion sentence pairs using a multitude of datasets, including Reddit Comments [18], S2ORC [19], Wiki Answers [20], PAQ [21],

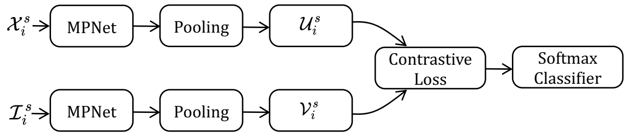


Fig. 3: Modified Sentence-BERT with MPNet.

MS MARCO [22], and GooAQ [23], among others. The MPNet is fine-tuned using a sentence pair similarity objective, which calculates cosine similarity between sentence pairs in a batch and compares it to true pairing using cross-entropy loss. We further fine-tune the MPNet for zero-shot intent detection task using a contrastive learning strategy.

Sentence-BERT. The design of Sentence-BERT is geared towards solving the computational overhead issues involved in sentence pair regression task, encountered while using BERT[13] and RoBERTa[24] like architectures. The Sentence-BERT overcomes this overhead by employing siamese and triplet loss [25]. In this work, we employ the Sentence-BERT model jointly with MPNet for our contrastive learning-based training. We integrate MPNet into Sentece-BERT for calculating sentence embeddings. Figure 3 depicts the employed Sentence-BERT architecture with integrated MPNet.

B. *Contrastive Learning.*

To facilitate robust zero-shot intent detection, we fine-tune the MPNet model with a contrastive objective. We leverage negative sampling to capture the utterance-to-intent correspondence. By randomly choosing K negative samples per positive sample, the model's goal is to learn the positive and negative association between utterance and intent pairs. We use the contrastive loss for guiding the training process. Our training approach provides the added advantage of steering the network away from learning the dataset-specific patterns and forces the model to learn more generic transferable associations between utterance and intent tokens. In the following, we provide further details about the training process.

Negative Sampling. Given a training example $(\mathcal{X}_i^s, \mathcal{I}_i^s) \in (X^s \times I^s)$, we sample K negative samples for $(\mathcal{X}_i^s, \mathcal{I}_i^s)$ denoted by $\bar{\mathcal{N}}_i^s = \{(\mathcal{X}_i^s, \bar{\mathcal{I}}_1^s), (\mathcal{X}_i^s, \bar{\mathcal{I}}_2^s), (\mathcal{X}_i^s, \bar{\mathcal{I}}_3^s), \ldots, (\mathcal{X}_i^s, \bar{\mathcal{I}}_K^s)\} \in X_i \times I_i \backslash \mathcal{I}_i^s$.

Input Embeddings. Let $\mathcal{F}_\theta$ denote our embedding model and $(\mathcal{X}_i^s, \mathcal{I}_i^s)$ be a training example. We get the embeddings $\mathcal{U}_i^s$ and $\mathcal{V}_i^s$ for $\mathcal{X}_i^s$ and $\mathcal{I}_i^s$, respectively, using:

$$\mathcal{U}_i^s = \mathcal{F}_\theta(\mathcal{X}_i^s) \tag{5}$$

$$\mathcal{V}_i^s = \mathcal{F}_\theta(\mathcal{I}_i^s) \tag{6}$$

where θ represents the model parameters.

Objective Function. We employ contrastive objective function for learning the sentence level embeddings. Consider that we are given a positive training example $(\mathcal{X}_i^s, \mathcal{I}_i^s)$ with a label of 1. Similarly, a negative training example $(\mathcal{X}_i^s, \bar{\mathcal{I}}_i^s)$ has a label 0. The contrastive learning objective is to pull the positive sample $(\mathcal{X}_i^s, \mathcal{I}_i^s)$ closer together in the embedding space, while the negative sample $(\mathcal{X}_i^s, \bar{\mathcal{I}}_i^s)$ is pushed away. Having calculated the embeddings for the training example $(\mathcal{X}_i^s, \mathcal{I}_i^s)$ as $\mathcal{U}_i^s$ and $\mathcal{V}_i^s$, the loss function is given below:

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathcal{U}_i^s.\mathcal{V}_i^s/\tau\right)}{\exp\left(\mathcal{U}_i^s.\bar{\mathcal{V}}_i^s/\tau\right)} \tag{7}$$

where $\bar{\mathcal{V}}_i^s$ represents the embedding of the negative sample's corresponding intent.

Generalization to New Unseen Intents. The ability to generalize to new unseen intents can be attributed to the contrastive objective of our proposed design. Our objective function is focused on learning the association between utterances and intent in such a way that utterance/intents pairs with high affinity are pulled together, while those not bearing much attraction are pushed away in the high dimensional manifold. Along the same lines, our approach lets the model learn a scoring function that assigns a score value between 0 and 1 to utterances/intent pairs. By doing that, the model's objective is refined towards learning generic features, rather than learning the dataset-specific distribution and their corresponding characteristics. Since the model's learned features are more token-oriented rather than dataset's distribution-oriented, the model learns the ability to distinguish the relationship between a nascent intent and corresponding utterance which may emerge in the future after the system's deployment.

C. *Implementation Detail*

Our model has a hidden size of 768 and the maximum allowed tokens are 384. We employ cosine similarity for evaluation between predicted and real labels. We use a learning rate of $2e^{-5}$ with *AdamW* optimizer. We trained our model for 1 epoch with a batch size of 4. Our optimization scheduler is *WarmupLinear* with 100 warmup steps. We only employ 200 update steps per epoch with a contrastive loss function. We trained our model on NVIDIA GeForce RTX 3090 24 GB graphics card. For few-shot experiments, we use 2000 update steps per epoch. We train for only 1 epoch. We use negative sampling for contrastive learning by using various ratios of positive to negative samples depending upon the cardinality of the dataset. For example, SNIPS has 7 classes, so we have the leverage of using 6 classes for composing negative pairs with 1 positive class. On similar lines, HWU64, BANKING77, and CLINC150 have a 1:20 negative sampling ratio for training. For our few-shot experiments, we employed a 1:2 positive-to-negative sampling ratio for SNIPS, HWU64, and BANKING77, whereas a 1:20 negative sampling ratio is used for CLINC150.

IV. Experimental Setup

For both zero-shot and few-shot intent detection tasks, we use the F1 score as our main evaluation metric.

A. *Datasets*

We employ four public benchmarks for intent detection. Table II presents important statistics about the datasets. (*i*) *CLINC150* [26] has 150 in-domain intent classes across 10 domains. Dataset also offers *out of scope* utterances.

TABLE I: Comparison in the zero-shot setting with the competing methods. The top row shows the test datasets. 2^{nd} row shows the training datasets. For example, when CLINC150, HWU64, and SNIPS are the training datasets, BANKING77 is the test dataset.

Test Dataset →	**BANKING77**			**CLINC150**			**HWU64**			**SNIPS**		
Method ↓ Train Dataset →	**CLINC150**	**HWU64**	**SNIPS**	**BANKING77**	**HWU64**	**SNIPS**	**CLINC150**	**BANKING77**	**SNIPS**	**CLINC150**	**BANKING77**	**HWU64**
DeViSE	0.001	0.001	0.04	0.001	0.002	0.062	0.005	0.001	0.086	0.002	0.001	0.002
Zero-shot DNN	0.509	0.509	0.509	0.587	0.587	0.587	0.426	0.426	0.426	0.682	0.682	0.682
albert-base-v2	0.001	0.002	0.231	0.504	0.493	0.310	0.001	0.001	0.272	0.669	0.553	0.751
albert-large-v2	0.001	0.001	0.013	0.512	0.001	0.001	0.001	0.001	0.341	0.035	0.035	0.035
roberta-base	0.488	0.323	0.273	0.001	0.001	0.376	0.001	0.441	0.297	0.737	0.504	0.744
roberta-large	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.035	0.050	0.035
bert-base-uncased	0.490	0.253	0.400	0.538	0.524	0.388	0.463	0.406	0.368	**0.801**	0.773	**0.857**
GPT2	0.047	0.035	0.005	0.261	0.088	0.010	0.138	0.126	0.059	0.340	0.29	0.162
T5	0.140	0.011	0.029	0.303	0.066	0.110	0.275	0.266	0.118	0.663	0.367	0.382
INTEND (This work)	**0.605**	**0.604**	**0.605**	**0.687**	**0.69**	**0.687**	**0.531**	**0.528**	**0.527**	0.795	**0.795**	0.797

TABLE II: Datasets statstics.

Dataset	**Samples**	**Vocab.**	**Avg. Length**	**Intents**
CLINC150	22.5K	7.2K	8.30	150
BANKING77	13K	4.5K	11.69	77
HWU64	11.1K	4.8K	6.56	64
SNIPS	14.4K	12.1K	9.00	7

(*ii*) *BANKING77* [27] contains 13083 utterances across 77 intents belonging to a single domain. (*iii*) *HWU64* [28] covers 21 domains with 64 intents. (*iv*) *SNIPS* [29] is a natural language understanding benchmark with 14.4k utterances spanned across 7 intents.

B. *Competing Models*

All the competing models have been trained and evaluated using the same setup as INTEND.

*1) **Zero-shot Methods**:* (*i*) *DeVise* [30] strives to find the compatibility between utterance/intent pair through a trainable linear projection. (*ii*) *Zero-shot DNN* [4] is designed to calculate the score between utterance and intent, and the highest affinity score is selected for prediction. (*iii*) *albert-base-v2* [31] is an MLM pre-trained on English. (*iv*) *albert-large-v2* [31] uses the same learning objectives as albert-base-v2. Contrasting differences are in configuration where albert-large-v2 employs 16 attention heads, 1024 hidden dimensions, and 24 repeating layers consisting of 17M parameters. (*v*) *roberta-base* [24] employs MLM objective and trained on 160 GB of text data. (*vi*) *roberta-large [24]* is the larger version of roberta-base trained with self-supervised learning and MLM objective on English corpus from various sources e.g., *BookCorpus, CC-News*. (*vii*) *bert-base-uncased* [13] is an uncased language model trained with MLM objective. It is trained on *BookCorpus* and *English Wikipedia*. Vocabulary size is 30,000 and all text is lowercased and tokenized with WordPiece. (*viii*) *GPT2* [32] employs causal language modeling (CLM) objective. GPT2 is trained in an autoregressive manner.(*ix*) t5-large [12] is an encoder-decoder architecture pre-trained on various text-to-text unsupervised and supervised tasks.

*2) **Few-shot Methods**:* (*i*) *GPT3* API from *OpenAI* has also been employed for few-shot intent detection comparison purposes. (*ii*) *DNNC (Discriminative nearest neighbor classification)* [9] with deep self-attention is also one of our baselines for few-shot intent detection comparison. In this work, BERT-style pairwise encoding is used to train a binary classifier for best-matched training example estimation for user input. (*iii*) *DialoGLUE (ConvBERT-DG)* [10] is a pre-trained model which is built on top of BERT [13] in order to mitigate its insufficiency when processing dialogue data. *ConvBERT* is trained on a large-scale open-domain dialogue corpus with 700 million conversations. *ConvBERT-DG* is a *ConvBERT*, but it is trained on full *DialoGLUE* [10] data in a supervised manner.

C. *Data Splits*

All the models are trained on the full training dataset S with all the utterance/intent pairs merged together from train/validation/test sets while keeping the test dataset T separate for evaluation. We split the test dataset T into validation and testing such that validation contains 25% of the data while the rest of the data is kept aside for testing. We employ stratified sampling for splitting validation and test sets.

V. Results

A. *Quantitative Analysis*

Comparison with Zero-shot Models. Table I presents the F1 score of all the models on a wide range of training and testing configurations. Our proposed INTEND outperforms all the competing models on all the configurations (with a few exceptions where it shows highly competitive performance) by a large margin. Specifically, when our testing dataset was BANKING77, INTEND is at least 10 percentage points more accurate than the second-best model(s) trained using CLINC150, HWU64, or SNIPS. Similar results can be observed when the evaluation datasets are CLINC150 and HWU64 where the minimum performance gap between INTEND and the second best model is 10 percentage points. We notice that all the baseline zero-shot intent detection models suffer to capture generic transferable patterns, resulting in poor F1 scores. We attribute this to the baseline models' tendency to overfit the training datasets. On the other hand, INTEND outperforms all the competing methods consistently and obtains a high F1 score. It is important to note that our proposed approach consistently demonstrates its ability to transfer the generic learned patterns to the test dataset irrespective of training datasets. It can be noticed that when the training dataset is CLINC150, HWU64, or SNIPS, and the evaluation dataset is BANKING77, INTEND's performance shows little to no variance that demonstrates its robustness to learn the transferable generic patterns. Similar observations can be noted for other experimental configurations. We argue that the variant of Sentence-BERT with MPNet in tandem trained

TABLE III: Comparison in the few-shot setting with competing methods. The results of baselines and our method are presented for different number (1-5) of training examples. Our model consistenly outperforms on 1-shot and 2-shot settings.

Dataset →	**CLINC150**			**BANKING77**			**HWU64**			**SNIPS**		
Shots ↓ Method →	DNNC	DialoGLUE (ConvBERT-DG)	INTEND	DNNC	DialoGLUE (ConvBERT-DG)	INTEND	DNNC	DialoGLUE (ConvBERT-DG)	INTEND	DNNC	DialoGLUE (ConvBERT-DG)	INTEND
1-shot	0.052	0.501	**0.818**	0.072	0.174	**0.671**	0.087	0.354	**0.613**	0.079	0.037	**0.904**
2-shot	0.689	0.793	**0.837**	0.514	0.438	**0.690**	0.469	0.632	**0.672**	0.053	0.037	**0.927**

with contrastive learning objective demonstrates the strong tendency of capturing generic dataset characteristics which are free from the underlying distribution of the training dataset. While on the other hand, we have several strong baseline models as well as pre-trained language models that follow state-of-the-art transformers architecture like autoencoding based, autoregressive based, and encoder/decoder (i.e., T5), but almost all the models suffer from a tendency to overfit to the training dataset when evaluated on a different test dataset. Similar performance gains for INTEND are observed in comparison with specialized zero-shot methods, where INTEND outperforms all the competing methods.

We also notice that the larger the model, the higher the overfitting tendency. For example, albert-large-v2 and roberta-large both show poor performance when employed for feature transfer from CLINC150 to BANKING77. SNIPS is a relatively simpler dataset with only 7 intents. It is also interesting to note that one of the baseline models (i.e., bert-base-uncased) shows good learning capability when the evaluation dataset is SNIPS, but at the same time, it fails to perform well on difficult datasets like BANKING77. Except for two configurations, our model outperforms all the models in all zero-shot experiments and demonstrates its zero-shot generalizable learning capabilities in comparison with strong baselines. In one of those settings, our model remains very competitive (i.e., 0.795 vs 0.801). To summarize the zero-shot setting experiments, we demonstrate that the exceptional generic learning capabilities of our model in a wide range of zero-shot experimental configurations.

Comparison with Few-shot Models. We also compared our model with the best-performing few-shot models, like [10] and [9]. We evaluate both the baselines and our proposed model on 1-shot and 2-shot settings. Table III summarizes the results for few-shot experiments. We report that our proposed model shows strong learning capability in 1-shot and 2-shot settings and consistently outperforms the competing baselines. When a few labeled training examples are available for training, our model remains competitive with the baselines. In fact, INTEND is still the best model among all the competitors, though the performance margin is generally smaller on challenging datasets (e.g., 4 percentage points for CLINC150 and HWU64). Keeping in view the true essence of zero-shot settings, our objective is to evaluate the various competing systems under "zero-shot" or "closer to zero-shot" settings. We argue that zero-shot models should achieve appreciable and consistent improvements with a minimum number of training shots. Focusing on 1-shot and 2-shot results, it can be noticed that our proposed model, when presented with 1-2 shots, achieves notable improvement in comparison with the system's zero-shot performances on that dataset. For example, let us focus on our model's performance on CLINC150 both in zero-shot and few-shot setups. In the zero-shot setting, our model manages to get an F1 score of 0.69 when pre-trained on HWU64. Under 1-shot setting, F1 score jumps from 0.69 to 0.818 which amounts to 18.5 percentage points improvement with just a single shot. Hence, we argue that our system remains true to the realistic zero-shot and few-shot settings, either with no data or bare minimum data.

TABLE IV: GPT3 API Results: F1 score on various datasets based upon single-shot prompt and 10-shot evaluation

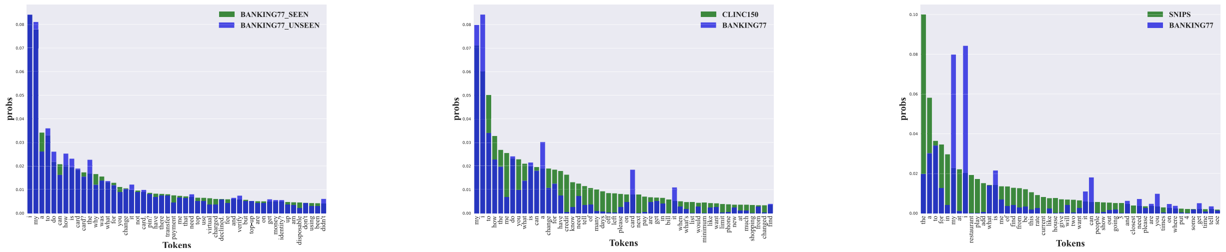
	BANKING77	**CLINC150**	**HWU64**	**SNIPS**
F1	0.51	0.59	0.43	0.68

Comparative Study with GPT-3. We also employ GPT-3 API for getting a comparative insight into our proposed strategy. We provide GPT-3 Q&A module with a prompt and corresponding *text* and *label* pairs. For example, we provided one utterance and label pair per class (i.e., equivalent of 1-shot setting) along with our designed prompt. We, then, tested GPT-3 with 10 utterances per class for each and recorded F1 score. Results are summarized in IV. Let's focus on BANKING77 results for comparison. F1 score achieved by GPT-3 on BANKING77 is 0.51 while INTEND manages 0.605 in zero-shot setting and 0.671 in the 1-shot experiment. Similarly, INTEND exhibits better performance on other benchmarks (see Table I and IV). The performance gains in case of INTEND can be attributed to our proposed contrastive learning objective.

B. *Qualitative Analysis*

It is a central requirement for zero-shot intent detection systems to accommodate new intents which may emerge in the future once the system is deployed. Our system is designed with this very requirement in mind, when scarcity of data becomes an issue and little to no training data is available. Similarly, closer to real-world zero-shot situations, our system demonstrates the ability to adapt to the requirement with minimum dependency upon the training data for accommodating new intents.

Distributional Divergence Analysis. We also present a comparative divergence analysis of training and testing datasets to demonstrate that the traditional zero-shot learning evaluation setup that splits a single dataset into seen and unseen sets to conducts experiments is not rigorous. Figure 4(a) presents the distributions of BANKING77 dataset when it is split into seen and unseen sets. We observe that the distributions of the seen (i.e., used for training) and unseen (i.e., used for evaluation) splits are very close to each other with a small KL divergence score of 0.13. We observed similar KL divergence scores when any dataset is split into seen and unseen sets. This result signifies that the training and testing distributions in the traditional zero-shot experimental setup are very close and do not allow for true zero-shot evaluations, when training and testing distributions are divergent. Figure 4(b) and 4(c)



(a) Seen/Unseen Splits from same dataset: BANKING77; *(KL=0.13)*.

(b) Distribution Plot for CLINC150 and BANKING77; *(KL=1.02)*.

(c) Distribution Plot for SNIPS and BANKING77; *(KL=1.38)*

Fig. 4: KL Divergence analysis highlights that if one dataset is split into seen and unseen sets (i.e., traditional zero-shot evaluation setting), the distributions for both sets show significantly low divergence. On the other hand, two different datasets generally have high KL divergence that facilitates rigorous (or true) zero-shot evaluations (best viewed in color).

present distributional plots for CLINC150 and BANKING77 (i.e., KL=1.02), and SNIPS and BANKING77 (i.e., KL=1.38). A comparatively, high KL divergence score can be noticed when the distributions of two different datasets are compared, thus facilitating true zero-shot evaluations. Based on our evaluation results, we notice that almost all the zero-shot intent models fail to perform well in this rigorous evaluation setting. However, our proposed method outperforms all the competing baselines in feature transfer task between two different datasets, i.e., when training and testing distributions are far from each other (see Table I). We argue that INTEND's better performance can be attributed to our contrastive learning design, in which we focus on the transferable generic feature learning.

We argue that highly divergent training and testing datasets are the most relevant and close to real-world scenarios for zero-shot intent detection systems. So, the zero-shot intent detection models should have the ability to accommodate such divergent intents. It can be noted that INTEND shows excellent feature transfer capability from SNIPS to CLINC150 *(F1=0.68)*. With this result, we believe that INTEND exhibits a strong tendency to accommodate newly emerging divergent intents. Similarly, INTEND exhibits a high F1 score (0.531) for HWU64 when trained on CLINC150. KL-divergence between CLINC150 and HWU64 is 1.14 which provides further evidence for high feature transfer capability of INTEND in case of newly emerging unseen intents.

VI. Related Work

Supervised Methods. An attention-based neural network for joint intent detection and slot filling is proposed in [1]. They also investigate different strategies to incorporate alignment information in the encoder-decoder framework and suggest introducing attention to alignment based RNN-models. A convolutional neural networks-based joint intent detection and slot-filling model is presented in [3]. They propose a neural version of the triangular CRF model for modeling intents and sequences jointly. The authors in [33] build upon the previous intent detection and slot-filling works, which consider these tasks separately. They propose to consider the cross-impact of these two tasks together and propose a bi-model-based RNN semantic frame parsing network and attempt to solve these two tasks jointly. Authors in [34] propose a non-autoregressive model for jointly solving intent detection and slot-filling tasks. In [35], authors propose to solve intent detection and slot filling with a stack propagation strategy for effectively using the intent information for slot filling task. All the works in the supervised setting require a large amount of labeled training data for each intent, whereas the focus of this work is to generalize to new unseen intent labels with no training data.

Zero-shot Methods. Zero-shot intent detection is of particular interest since it tackles mitigating the data scarcity issue. The usage of label ontologies [36], [37], external knowledge sources [6], outlier detection algorithm [38], [39], among others, have been explored to facilitate zero-shot intent detection. Moreover, mapping of intent and utterances to the same embedding space [30], [4] and capsule neural networks [40] have also been investigated. While these works show promising results in the zero-shot setting, they evaluate the models by splitting a single dataset into training and testing datasets – not a rigorous experimental setup. In this work, we focus on the more practically relevant and thorough zero-shot evaluations when the distributions of seen and unseen intents are divergent.

Conclusion

We present a robust zero-shot intent detection strategy. Our approach focuses on generic feature learning by employing contrastive learning and leveraging the power of pre-trained language models. We use a variant of Sentence-BERT which is equipped with modern MPNet architecture. Prior works split a single training dataset into seen and unseen intents that may lead to misclassifications when the model is presented with divergent emerging intents in the functional lifespan of the system. We suggest keeping the training and testing datasets totally different, having different distributions facilitates robust zero-shot evaluations. Our proposed approach accommodates new unseen divergent intents flawlessly.

References

[1] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," *arXiv preprint arXiv:1609.01454*, 2016.

[2] C. Zhang, Y. Li, N. Du, W. Fan, and P. S. Yu, "Joint slot filling and intent detection via capsule neural networks," *arXiv preprint arXiv:1812.09471*, 2018.

[3] P. Xu and R. Sarikaya, "Convolutional neural network based triangular crf for joint intent detection and slot filling," in *2013 ieee workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 78–83.

[4] A. Kumar, P. R. Muddireddy, M. Dreyer, and B. Hoffmeister, "Zero-shot learning across heterogeneous overlapping domains," in *INTERSPEECH*, 2017.

[5] A. Siddique, F. Jamour, and V. Hristidis, "Linguistically-enriched and context-awarezero-shot slot filling," in *Proceedings of the Web Conference 2021*, 2021, pp. 3279–3290.

[6] A. Siddique, F. Jamour, L. Xu, and V. Hristidis, "Generalized zero-shot intent detection via commonsense knowledge," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1925–1929.

[7] A. Siddique, M. Maqbool, K. Taywade, and H. Foroosh, "Personalizing task-oriented dialog systems via zero-shot generalizable reward function," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1787–1797.

[8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[9] J. Zhang, K. Hashimoto, W. Liu, C.-S. Wu, Y. Wan, P. Yu, R. Socher, and C. Xiong, "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)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5064–5082. [Online]. Available: https://aclanthology.org/2020.emnlp-main.411

[10] S. Mehri, M. Eric, and D. Hakkani-Tur, "Dialoglue: A natural language understanding benchmark for task-oriented dialogue," *arXiv preprint arXiv:2009.13570*, 2020.

[11] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mass: Masked sequence to sequence pre-training for language generation," *arXiv preprint arXiv:1905.02450*, 2019.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[15] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 857–16 867, 2020.

[16] A. Siddique, S. Oymak, and V. Hristidis, "Unsupervised paraphrasing via deep reinforcement learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1800–1809.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] M. Henderson, P. Budzianowski, I. Casanueva, S. Coope, D. Gerz, G. Kumar, N. Mrkšić, G. Spithourakis, P.-H. Su, I. Vulić *et al.*, "A repository of conversational datasets," *arXiv preprint arXiv:1904.06472*, 2019.

[19] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld, "S2ORC: The semantic scholar open research corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4969–4983. [Online]. Available: https://aclanthology.org/2020.acl-main.447

[20] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1156–1165.

[21] P. Lewis, Y. Wu, L. Liu, P. Minervini, H. Küttler, A. Piktus, P. Stenetorp, and S. Riedel, "Paq: 65 million probably-asked questions and what you can do with them," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1098–1115, 2021.

[22] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "Ms marco: A human generated machine reading comprehension dataset," in *CoCo@ NIPS*, 2016.

[23] D. Khashabi, A. Ng, T. Khot, A. Sabharwal, H. Hajishirzi, and C. Callison-Burch, "Gooaq: Open question answering with diverse answer types," *arXiv preprint arXiv:2104.08727*, 2021.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[26] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1311–1316. [Online]. Available: https://aclanthology.org/D19-1131

[27] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Online: Association for Computational Linguistics, Jul. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.nlp4convai-1.5

[28] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, "Benchmarking natural language understanding services for building conversational agents," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*. Springer, 2021, pp. 165–183.

[29] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[30] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[31] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[33] Y. Wang, Y. Shen, and H. Jin, "A bi-model based rnn semantic frame parsing model for intent detection and slot filling," *arXiv preprint arXiv:1812.10235*, 2018.

[34] D. Wu, L. Ding, F. Lu, and J. Xie, "Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling," *arXiv preprint arXiv:2010.02693*, 2020.

[35] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," *arXiv preprint arXiv:1909.02188*, 2019.

[36] M. Yazdani and J. Henderson, "A model of zero-shot learning of spoken language understanding," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 244–249.

[37] E. Ferreira, B. Jabaian, and F. Lefevre, "Online adaptative zero-shot learning spoken language understanding using word-embedding," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5321–5325.

[38] V. Gangal, A. Arora, A. Einolghozati, and S. Gupta, "Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7764–7771.

[39] G. Yan, L. Fan, Q. Li, H. Liu, X. Zhang, X.-M. Wu, and A. Y. Lam, "Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 1050–1060.

[40] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," *arXiv preprint arXiv:1809.00385*, 2018.