Generalized Zero-shot Intent Detection via Commonsense Knowledge

A.B. Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis
msidd005,fuadj,lxu051{@ucr.edu},vagelis@cs.ucr.edu
University of California, Riverside
Riverside, CA, USA

ABSTRACT

Identifying user intents from natural language utterances is a crucial step in conversational systems that has been extensively studied as a supervised classification problem. However, in practice, new intents emerge after deploying an intent detection model. Thus, these models should seamlessly adapt and classify utterances with both seen and unseen intents – unseen intents emerge after deployment and they do not have training data. The few existing models that target this setting rely heavily on the training data of seen intents and consequently overfit to these intents, resulting in a bias to misclassify utterances with unseen intents into seen ones. We propose RIDE: an intent detection model that leverages commonsense knowledge in an unsupervised fashion to overcome the issue of training data scarcity. RIDE computes robust and generalizable *relationship meta-features* that capture deep semantic relationships between utterances and intent labels; these features are computed by considering how the concepts in an utterance are linked to those in an intent label via commonsense knowledge. Our extensive experimental analysis on three widely-used intent detection benchmarks shows that relationship meta-features significantly improve the detection of both seen and unseen intents and that RIDE outperforms the state-of-the-art models.

CCS CONCEPTS

• **Computing methodologies → Feature selection**; *Discourse, dialogue and pragmatics*; *Lexical semantics.*

KEYWORDS

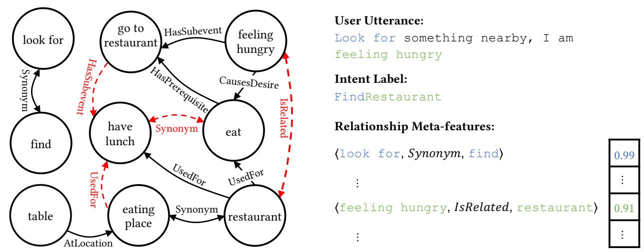
generalized zero-shot learning, out of domain intent detection, natural language understanding, natural language processing.

ACM Reference Format:

A.B. Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized Zero-shot Intent Detection via Commonsense Knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3462985

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07.
https://doi.org/10.1145/3404835.3462985



(a) Automatic Commonsense KG Completion (b) Relationship Meta-feature Generation

Figure 1: (a) Link Predictor learns from KG (solid lines) to predict missing edges (dashed lines), (b) Example utterance, intent, and computation of relationship meta-features that facilitate GZS intent detection.

1 INTRODUCTION

Virtual assistants such as Amazon Alexa and Google Assistant allow users to perform a variety of tasks (e.g., Alexa Skills) through a natural language interface. For example, a user can set an alarm by simply issuing the utterance *"Wake me up tomorrow at 10 AM"* to a virtual assistant, and the assistant is expected to understand that the user's intent (i.e., "AddAlarm") is to invoke the alarm module, then set the requested alarm accordingly. Intent detection is typically the first step towards performing any task in conversational systems and it is a challenging problem due to the vast diversity in user utterances. The challenge is further exacerbated in the more practically relevant setting where intents are added over time. This setting is an instance of the *generalized zero-shot classification problem* [10]: labeled training utterances are available only for seen intents but are unavailable for unseen ones, and at inference time, models do not have prior knowledge on whether the utterances they receive imply seen or unseen intents. This setting is the focus of this paper.

Little research has been conducted on building generalized zero-shot (GZS) models for intent detection, with little success. Earlier works [4, 18, 31] used zero-shot (ZS) learning to train an intent classification model that could classify utterances from unseen intent classes through transferring knowledge from seen classes. The test set in the standard ZS setting is not representative of the real world, as it exclusively includes samples from unseen classes (as opposed to having samples from both seen and unseen classes as in the GZS setting). ZS methods perform poorly in the GZS setting [3, 27], which is primarily caused by their strong bias towards seen classes; ZS intent detection models misclassify almost all test samples from unseen classes into seen ones [22, 37, 38].

To mitigate the issue of training data scarcity for unseen intents and ZS models' inability to effectively handle the GZS setting, we propose incorporating commonsense knowledge into a

GZS intent detection model. We argue that such knowledge, if incorporated properly, helps overcome training data scarcity and allows detecting intents regardless of whether they are seen or not, given that commonsense knowledge is uniform across intents. We leverage ConceptNet [32] — a rich and widely-used commonsense knowledge graph (KG) that captures a large subset of knowledge in a semi-structured format (i.e., facts in the form ⟨`head`, *relation*, `tail`⟩ such as ⟨`apple`, *IsA*, `fruit`⟩). Given that ConceptNet is incomplete, similarly to other KGs, we pre-train a link predictor [16] that learns from an existing KG to infer novel edges (i.e., relationships) among nodes (i.e., head/tail) to overcome the missing information challenge. Figure 1 (a) presents a toy commonsense KG where a link predictor can learn to infer missing facts such as ⟨`feeling hungry`, *IsRelated*, `restaurant`⟩ from existing facts such as ⟨`feeling hungry`, *CausesDesire*, `eat`⟩ and ⟨`restaurant`, *UsedFor*, `eat`⟩. We infuse the knowledge from our link predictor into our model by extracting *relationship meta-features*. These features quantify the level of relevance between an utterance and an intent in the form of relationship weights, where each weight describes the level of relatedness between the phrases in an utterance and an intent label based on a certain relationship type. Figure 1 (b) shows an example utterance, an intent label, and inference about the relationships between the phrases in the utterance and the intent label in the from of a relationship meta-features. Relationship meta-features augment embeddings using commonsense knowledge, which significantly reduces our model's reliance on the scarcely available seen intents training data. Furthermore, these features reduce our model's bias towards seen intents given that they are similarly computed for both seen and unseen intents; i.e., they are domain-oblivious.

Our model, RIDE[1,2], combines relationship meta-features with contextual word embeddings [24], and feeds the combined feature vectors into a trainable prediction function. RIDE is able to accurately detect both seen and unseen intents in utterances. Our extensive experimental analysis using the three widely used benchmarks SNIPS [6], SGD [25], and MultiWOZ [42] show that our model outperforms the state-of-the-art (SOTA) model in F1 scores on unseen intents in the GZS setting by at least 25.66%.

A secondary contribution of this paper is that we managed to further improve the performance of GZS intent detection by employing Positive-Unlabeled (PU) learning [8] to predict if a new utterance belongs to a seen or unseen intent. PU learning assists intent detection models by mitigating their bias towards classifying most utterances into seen intents. A PU classifier is able to perform binary classification after being trained using only positive and unlabeled examples. We found out that the PU classifier also improves the performance of existing intent detection works. Our model, however, outperforms existing ones regardless of the PU classifier's integration.

2 PRELIMINARIES

GZS Intent Detection. Let $\mathcal{S} = \{\mathcal{I}_1, \cdots, \mathcal{I}_k\}$ be a set of seen intents and $\mathcal{U} = \{\mathcal{I}_{k+1}, \cdots, \mathcal{I}_n\}$ be a set of unseen intents where $\mathcal{S} \cap \mathcal{U} = \varnothing$. Let $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_m\}$ be a set of labeled training utterances where each training utterance $\mathcal{X}_i \in \mathcal{X}$ is described with a tuple $(\mathcal{X}_i, \mathcal{I}_j)$ such that $\mathcal{I}_j \in \mathcal{S}$. An intent $\mathcal{I}_j$ is comprised of an

Algorithm 1: RMG

Input: $\mathcal{R} = \{r_1, \cdots, r_t\}$: relations in KG
$\mathcal{G}_i = \{g_1, \cdots, g_q\}$: utterance n-grams
$\mathcal{I}_j = \{\mathcal{A}, \mathcal{O}\}$: intent's Action and Object

Output: $\mathbf{e}_{relationship}$: $\mathcal{X}_i$-$\mathcal{I}_j$ relationship meta-features

Let $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{A}}} = \mathrm{RM}\,(\mathcal{A}, \mathcal{G}_i, \longrightarrow)$ // Action to utterance
Let $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{O}}} = \mathrm{RM}\,(\mathcal{O}, \mathcal{G}_i, \longrightarrow)$ // Object to utterance
Let $\mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{A}}} = \mathrm{RM}\,(\mathcal{A}, \mathcal{G}_i, \longleftarrow)$ // utterance to Action
Let $\mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{O}}} = \mathrm{RM}\,(\mathcal{O}, \mathcal{G}_i, \longleftarrow)$ // utterance to Object
Let $\mathbf{e}_{relationship} = [\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{A}}}, \mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{O}}}, \mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{A}}}, \mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{O}}}]$
return $\mathbf{e}_{relationship}$

Function RM($concept, phrases, direction$):
  Let $\mathbf{e}$ = []
  foreach $r \in \mathcal{R}$ **do**
    if $direction = \longrightarrow$ **then**
      Let $p = \mathrm{Max}\,(LP(concept, r, g))$ for $g \in phrases$
    if $direction = \longleftarrow$ **then**
      Let $p = \mathrm{Max}\,(LP(g, r, concept))$ for $g \in phrases$
    e.append(p)
  return e

Action and an *Object* and takes the form "ActionObject"[3] (e.g., "FindRestaurant"); an Action describes a user's request or activity and an Object describes the entity pointed to by an Action [5, 33, 34]. Given a test utterance $\mathcal{X}_i'$, the problem is to predict a label $\mathcal{I}_j' \in \mathcal{S} \cup \mathcal{U}$.

Link Prediction in Knowledge Graphs. We pre-train a SOTA link prediction model (LP) [16] on ConceptNet [32] to score novel facts that are not necessarily present in the knowledge graph. Given a triple (i.e., fact) in the form ⟨`head`, *relation*, `tail`⟩, a link prediction model scores the triple with a value between 0 and 1, which quantifies the level of validity of the given triple.

Positive-Unlabeled Learning Positive-Unlabeled (PU) classifiers learn a standard binary classifier in the unconventional setting where labeled negative training examples are unavailable. The PU classifier [8] learns a probabilistic function $f(\mathcal{X}_i)$ that estimates $P(\mathcal{I}_j \in \mathcal{S} \mid \mathcal{X}_i)$ as closely as possible. We train a PU classifier using our training set (utterances with only seen intents labeled as positive) and validation set (utterances with both seen and unseen intents as unlabeled). We use 512-dimensions sentence embedding as features when using the PU classifier, generated using a pre-trained universal sentence encoder [2].

3 OUR APPROACH

Given an input utterance $\mathcal{X}_i$, our model first invokes the PU classifier (if it is available) to predict whether $\mathcal{X}_i$'s intent belongs to set $\mathcal{S}$ or $\mathcal{U}$. Then, relationship meta-features, utterance embedding, and intent embedding are concatenated and fed into a trainable prediction function that predicts the probability $P(\mathcal{I}_j|\mathcal{X}_i) \in [0, 1]$. Finally, our model outputs the intent with the highest compatibility probability, i.e., $\mathrm{argmax}_{\mathcal{I}_j} P(\mathcal{I}_j|\mathcal{X}_i)$.

Relationship meta-features generator (RMG). It extracts features by utilizing the "ActionObject" structure of intent labels

[1] RIDE: Relationship Meta-features Assisted Intent DEtection.
[2] https://github.com/RIDE-SIGIR/GZS

[3] If intents are described using a complex textual description, Actions and Objects can be extracted using existing NLP tools such as dependency parsers.

Table 1: Dataset statistics.

Dataset	# of Samples	Vocab. Size	Avg. Length	# of Intents
SNIPS	14.2K	10.8K	9.05	7
SGD	57.2K	8.8K	10.62	46
MultiWOZ	30.0K	9.7K	11.07	11

and commonsense KG. RMG takes the following inputs: a set of relations in a knowledge graph (35 in the case of ConceptNet) $\mathcal{R} = \{r_1, r_2, .., r_t\}$; the set of n-grams $\mathcal{G}_i = \{g_1, g_2, .., g_q\}$ that correspond to the input utterance $\mathcal{X}_i$, where $|\mathcal{G}| = q$; and an intent label $\mathcal{I}_j = \{\mathcal{A}, \mathcal{O}\}$, where $\mathcal{A}$ and $\mathcal{O}$ are the Action and Object components of the intent, respectively. RMG computes a relationship meta-features vector in four steps, where each step results in a vector of size $|\mathcal{R}|$. The smaller vectors are: $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{A}}}$, $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{O}}}$, $\mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{A}}}$, and $\mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{O}}}$, where $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{A}}}$ captures the weights of Action to utterance relationships and $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{O}}}$ captures the weights of Object to utterance relationships. The remaining two vectors capture relationship weights in the other direction; i.e., utterance to Action/Object, respectively. Capturing bi-directional relationships is important because a relationship in one direction does not necessarily imply one in the other direction. The final output of RMG is the concatenation of the four aforementioned vectors.

RMG computes $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{A}}}$ by considering the strength of each relation in $\mathcal{R}$ between $\mathcal{A}$ and each n-gram in $\mathcal{G}_i$. That is, $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{A}}}$ has $|\mathcal{R}|$ cells, where each cell corresponds to a relation $r \in \mathcal{R}$. Each cell is computed by taking $max(LP(\mathcal{A}, r, g))$ over all $g \in \mathcal{G}_i$. $LP(\texttt{head}, relation, \texttt{tail})$ outputs the probability that the fact represented by the triple $\langle \texttt{head}, relation, \texttt{tail} \rangle$ exists. The vector $\mathbf{e}_{\mathcal{X}_i}^{\overrightarrow{\mathcal{O}}}$ is computed similarly, but with passing $\mathcal{O}$ instead of $\mathcal{A}$ when invoking the link predictor. The vectors $\mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{A}}}$ and $\mathbf{e}_{\mathcal{X}_i}^{\overleftarrow{\mathcal{O}}}$ are computed similarly, but with swapping the head and tail when invoking the link predictor. Algorithm 1 outlines the previous process. Finally, the meta-features are passed through a linear layer with sigmoid activation for normalization.

Utterance and Intent Encoders. We use bi-directional LSTM to produce a d-dimensional representation of the given utterance $\mathcal{X}_i = \{w_1, w_2, \cdots, w_u\}$ with u words, where contextual embeddings from a pre-trained ELMo model and POS tags are employed to embed each word. We encode intent labels similarly to produce an intent embedding $\mathbf{e}_{intent} \in \mathbb{R}^d$.

Training. The training examples are of the form $((\mathcal{X}_i, \mathcal{I}_j), \mathcal{Y})$, where $\mathcal{Y}$ is a binary label representing whether the utterance-intent pair $(\mathcal{X}_i, \mathcal{I}_j)$ are compatible or not. We prepare our training data by assigning a label of 1 to the available utterance-intent pairs (where intents are seen ones); these constitute positive training examples. We create a negative training example for each positive one by corrupting the example's intent by modifying their Action, Object, or both. We train the model by minimizing the cross-entropy loss over all the training examples.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We used three popular intent detection benchmarks: SNIPS, SGD, and MultiWOZ; Table 1 summarizes the statistics of these datasets. SNIPS [6] is a crowd-sourced single-turn NLU benchmark with 7 intents across different domains. SGD [25] is a comprehensive and challenging dataset with 46 intents across 16 domains. MultiWOZ [42] is a well-known dataset which has utterances that span 11 intents

Evaluation Methodology. We use F1 scores and report the per class averages weighted by the respective class support.

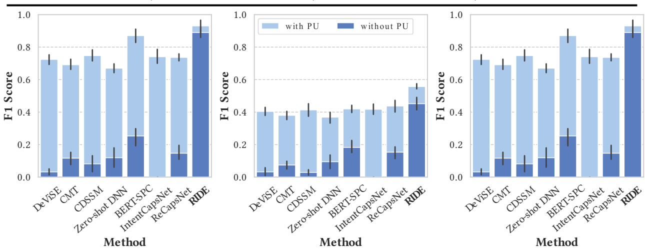
Dataset splits. All models are trained on a subset of utterances implying seen intents. At inference time, test utterances are drawn from a set that contains utterances implying a mix of seen and unseen intents (disjoint set from the training set). We decided the train/test splits for each dataset as follows: For SNIPS, we first randomly selected 5 out of 7 intents and designated them as seen intents. We then selected 70% of the utterances that imply any of the 5 seen intents for training. The test set consists of the remaining 30% utterances in addition to all utterances that imply any of the 2 unseen intents. For SGD, we used the standard splits proposed by the dataset authors. Specifically, the test set includes utterances that imply 8 unseen intents and 26 seen intents. For MultiWOZ, we used 70% of the utterances that imply 8 (out of 11) randomly selected intents for training and the rest of the utterances for testing; we report average results over 10 runs for all the datasets.

Competing Methods. We compare against the following models:
SEG [40]: semantic-enhanced gaussian mixture model coupled with a density-based outlier detection algorithm LOF.
ReCapsNet-ZS [21]: employs a capsule neural network (CapsNet) and a dimensional attention module to learn generalizable transformational metrices from seen intents.
IntentCapsNet [36]: utilizes CapsNet and routing-by-agreement to adapt to unseen intents. This model was originally proposed for detecting intents in the standard ZS setting, so we extended it to support the GZS setting with the help of its authors.
Other Baseline Models. (*i*) Zero-shot DDN [18]: achieves zero-shot capabilities by projecting utterances and intent labels into the same high-dimensional embedding space. (*ii*) CDSSM [4]: utilizes a convolutional deep structured semantic model to generate embeddings for unseen intents. (*iii*) CMT [31]: employs non-linearity in the compatibility function between utterances and intents to find the most compatible unseen intents. (*iv*) DeViSE [12]: originally proposed for zero-shot image classification that learns a linear compatibility function. (*v*) BERT Sentence Pair Classifier (BERT-SPC): BERT is pre-trained on the sentence-pair classification task and fine-tuned on the utterances from the seen intents. *Note that we extended the baseline ZS models to support the GZS setting.*

Implementation Details. We trained our link predictor on the lemmatized version of ConceptNet KG (1 million nodes, 2.7 million edges, and 35 relation types). The link predictor has two 200-dimensional embedding layers and a negative sampling ratio of 10; it is trained for $1,000$ epochs using Adam optimizer with a learning rate of 0.05, L2 regularization value of 0.1, and batch size of 4800. Our relationship meta-features generator takes in an utterance's n-grams with n $\leq$ 4 and an intent label, and it uses the pre-trained link predictor to produce relationship meta-features with 140 dimensions. Our utterance and intent encoders use pre-trained ELMo contextual word embeddings with 1024 dimension and POS tags embeddings with 300 dimension, and two-layer bidirectional LSTM with 300 dimensions. Our prediction function has two dense layers

Table 2: Main results: F1 scores for competing models.

Method	SNIPS		SGD		MultiWOZ	
	Unseen	Seen	Unseen	Seen	Unseen	Seen
DeViSE	0.0439	0.6521	0.0177	0.5451	0.0270	0.5770
CMT	0.0910	0.6639	0.0621	0.5803	0.0679	0.6216
CDSSM	0.0484	0.7028	0.0284	0.6379	0.0244	0.6515
Zero-shot DNN	0.1273	0.6687	0.1168	0.6098	0.1149	0.6012
BERT-SPC	0.2761	0.7152	0.1872	0.6401	0.1932	0.6413
IntentCapsNet	0.0000	0.6532	0.0000	0.5508	0.0000	0.6038
ReCapsNet	0.1601	0.6783	0.1331	0.5751	0.1467	0.6170
SEG	0.6991	0.8651	0.4032	0.6356	0.4143	0.6456
RIDE w/o PU	0.9103	0.8799	0.4634	0.8295	0.4645	0.8816
RIDE $/w$ PU	**0.9254**	**0.9080**	**0.5734**	**0.8298**	**0.5206**	**0.8847**



(a) SNIPS dataset **(b) SGD dataset** **(c) MultiWOZ dataset**

Figure 2: F1 scores for unseen intents for the competing models after integrating a PU classifier.

with relu and softmax activation. Our model is trained for up to 200 epochs with early stopping using Adam optimizer and a cross entropy loss with initial learning rate of 0.001 and ReduceLROnPlateau scheduler with 20 patience epochs. It uses a dropout rate of 0.3,a batch size of 32 and a negative sampling ratio of up to 6.

4.2 Results

Main Results. Table 2 shows F1 scores for all competing models. For both seen and unseen intents, our model RIDE outperforms all other competing models with a large margin. Specifically, RIDE achieves 32.37%, 42.21%, and 25.66% better F1 scores than the SOTA model SEG on SNIPS, SGD, and MultiWOZ for unseen intents, respectively. Moreover, our model consistently achieves the highest F1 score on seen intents, which confirms its generalizability. We highlight that RIDE outperforms SEG regardless of whether a PU classifier is incorporated or not. For SNIPS, the role of the PU classifier is negligible. Whereas, for SGD and MultiWOZ, which are more challenging datasets, the PU classifier causes significant improvements in F1 scores.

Effect of PU Classifier on Other Models. Figure 2 presents F1 scores of all models with and without PU classifier. A PU classifier significantly improves the results of all the competing models. For instance, the IntentCapsNet model with a PU classifier achieves an F1 score of 74% for unseen intents on SNIPS dataset compared to an F1 score of less than 0.01% without the PU classifier. Interestingly, our model RIDE without PU classifier outperforms all the competing models even when a PU classifier is incorporated into them, which highlights that the PU classifier is not the component that does the heavy lifting in our model.

Ablation Study. We present the results of our ablation study in Table 3. Utilizing utterance and intent embeddings only (i.e., UI-Embed) results in very low F1 score, i.e., 23.67% on SNIPS. Employing relationship meta-features only (i.e., Rel-M) results in significantly better results: an F1 score of 71.03% on SNIPS. When

Table 3: Ablation study: F1 scores for unseen intents.

Configuration	SNIPS	SGD	MultiWOZ
UI-Embed w/o PU	0.2367	0.1578	0.1723
Rel-M w/o PU	0.7103	0.3593	0.3321
RIDE w/o PU	0.9103	0.4634	0.4645
UI-Embed $/w$ PU	0.7245	0.4202	0.4124
Rel-M $/w$ PU	0.8463	0.5167	0.4781
RIDE $/w$ PU	**0.9254**	**0.5734**	**0.5206**

utterance and intent embeddings are used in conjunction with relationship meta-features (i.e., RIDE w/o PU), it achieves a better F1 score compared to the Rel-M or UI-Embed configurations. A similar trend can be observed for the other datasets as well. Finally, when our entire model is deployed (i.e., RIDE $/w$ PU), it achieves the best results on all datasets.

5 RELATED WORK

The deep neural networks have proved highly effective for many critical NLP tasks [9, 15, 20, 23, 28, 30, 35, 43] including intent detection. Supervised intent detection works [17, 20, 26, 39, 43] assume the availability of a large amount of labeled training data for all intents to learn discriminative features. Whereas standard zero-shot intent detection models [1, 7, 11, 13, 18, 35, 41] assume that all utterances faced at inference time imply unseen intents only. Extending such works to handle the generalized zero-shot intent detection setting (i.e., removing the aforementioned assumptions) is nontrivial. The authors in [21] attempted to accommodate GZS setting by adding a dimensional attention module to a capsule network that learns generalizable transformation matrices from seen intents. Recently, the authors in [40] proposed using a density-based outlier detection algorithm LOF [1] and SEG mixture model with large margin loss to learn class-concentrated embeddings to detect unseen intents. In contrast, we leverage a rich commonsense knowledge graph to capture deep semantic and discriminative relationships between utterances and intents, which significantly reduces the bias towards classifying unseen intents into seen ones. In a related, but orthogonal, line of research, the authors in [14, 19, 23] addressed intent detection in the context of dialog state tracking where an annotated dialog state and conversation history are available in addition to an input utterance. In contrast, this work and the SOTA models we compare against in our experiments only consider an utterance without having access to any dialog state elements. Interested readers can refer to [29] for more details about this work.

6 CONCLUSION

We have presented an accurate generalized zero-shot intent detection model. Our extensive experimental analysis on three intent detection benchmarks show that our model achieves 25.66% to 42.21% better F1 score than the SOTA model for unseen intents. The main novelty of our model is its utilization of relationship meta-features and limited reliance on training data. Furthermore, our idea of integrating Positive-Unlabeled learning in GZS intent detection models further improves our models' performance, and significantly improves the accuracy of existing models as well.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation (NSF) under grants IIS-1838222 and IIS-1901379.

REFERENCES

[1] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 93–104.

[2] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*. Springer, 52–68.

[4] Yun-Nung Chen, Dilek Hakkani-Tür, and Xiaodong He. 2016. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6045–6049.

[5] Zhiyuan Chen, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Identifying intention posts in discussion forums. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1041–1050.

[6] 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* (2018).

[7] Yann N Dauphin, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2013. Zero-shot learning for semantic utterance classification. *arXiv preprint arXiv:1401.0509* (2013).

[8] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 213–220.

[9] Umar Farooq, A. B. Siddique, Fuad Jamour, Zhijia Zhao, and Vagelis Hristidis. 2020. App-Aware Response Synthesis for User Reviews. In *2020 IEEE International Conference on Big Data (Big Data)*. 699–708.

[10] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 21–37.

[11] Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefevre. 2015. Online adaptative zero-shot learning spoken language understanding using word-embedding. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5321–5325.

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[13] Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2019. Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection In Task Oriented Dialog. *arXiv preprint arXiv:1912.12800* (2019).

[14] Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev. 2020. Goal-oriented multi-task bert-based dialogue state tracker. *arXiv preprint arXiv:2002.02450* (2020).

[15] Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, Fast, Accurate Intent Classification and Slot Labeling for Goal-Oriented Dialogue Systems. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. 46–55.

[16] Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in neural information processing systems*. 4284–4295.

[17] Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016. Intent detection using semantically enriched word embeddings. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 414–419.

[18] Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Björn Hoffmeister. 2017. Zero-Shot Learning Across Heterogeneous Overlapping Domains.. In *INTERSPEECH*. 2914–2918.

[19] Miao Li, Haoqi Xiong, and Yunbo Cao. 2020. The sppd system for schema guided dialogue state tracking challenge. *arXiv preprint arXiv:2006.09035* (2020).

[20] Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454* (2016).

[21] Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert YS Lam. 2019. Reconstructing Capsule Networks for Zero-shot Intent Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4801–4811.

[22] Kun Liu, Wu Liu, Huadong Ma, Wenbing Huang, and Xiongxiong Dong. 2019. Generalized zero-shot learning for action recognition with web-scale video data. *World Wide Web* 22, 2 (2019), 807–824.

[23] Yue Ma, Zengfeng Zeng, Dawei Zhu, Xuan Li, Yiying Yang, Xiaoyuan Yao, Kaijie Zhou, and Jianping Shen. 2019. An end-to-end dialogue state tracking system with machine reading comprehension and wide & deep classification. *arXiv preprint arXiv:1912.09297* (2019).

[24] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[25] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *arXiv preprint arXiv:1909.05855* (2019).

[26] Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[27] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. 2012. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 7 (2012), 1757–1772.

[28] A.B. Siddique, Fuad Jamour, and Vagelis Hristidis. 2021. Linguistically-Enriched and Context-Aware Zero-shot Slot Filling. In *Proceedings of the Web Conference 2021*. Association for Computing Machinery, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449870

[29] AB Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized Zero-shot Intent Detection via Commonsense Knowledge. *arXiv preprint arXiv:2102.02925* (2021).

[30] A. B. Siddique, Samet Oymak, and Vagelis Hristidis. 2020. Unsupervised Paraphrasing via Deep Reinforcement Learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1800–1809. https://doi.org/10.1145/3394486.3403231

[31] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*. 935–943.

[32] Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975* (2016).

[33] Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open Intent Extraction from Natural Language Interactions. In *Proceedings of The Web Conference 2020*. 2009–2020.

[34] Jinpeng Wang, Gao Cong, Wayne Xin Zhao, and Xiaoming Li. 2015. Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 318–324.

[35] Kyle Williams. 2019. Zero Shot Intent Classification Using Long-Short Term Memory Networks. *Proc. Interspeech 2019* (2019), 844–848.

[36] Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2018. Zero-shot user intent detection via capsule neural networks. *arXiv preprint arXiv:1809.00385* (2018).

[37] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence* 41, 9 (2018), 2251–2265.

[38] Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4582–4591.

[39] Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*. IEEE, 78–83.

[40] Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS Lam. 2020. 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*. 1050–1060.

[41] Majid Yazdani and James Henderson. 2015. A model of zero-shot learning of spoken language understanding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 244–249.

[42] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online, 109–117. https://doi.org/10.18653/v1/2020.nlp4convai-1.13

[43] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2018. Joint slot filling and intent detection via capsule neural networks. *arXiv preprint arXiv:1812.09471* (2018).