Unknown Intent Detection Using Gaussian Mixture Model with an Application to Zero-shot Intent Classification

Lu Fan[1*] **Guangfeng Yan**[1*] **Qimai Li**[1*]
Han Liu[1] **Xiaotong Zhang**[1] **Albert Y.S. Lam**[2] Xiao-Ming Wu[1†]
Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.[1]
Fano Labs, Hong Kong S.A.R.[2]
{csgfyan, cslfan, csqmli}@comp.polyu.edu.hk
{cshliu, csxtzhang, csxmwu}@comp.polyu.edu.hk, albert@fano.ai

Abstract

User intent classification plays a vital role in dialogue systems. Since user intent may frequently change over time in many realistic scenarios, unknown (new) intent detection has become an essential problem, where the study has just begun. This paper proposes a semantic-enhanced Gaussian mixture model (SEG) for unknown intent detection. In particular, we model utterance embeddings with a Gaussian mixture distribution and inject dynamic class semantic information into Gaussian means, which enables learning more class-concentrated embeddings that help to facilitate downstream outlier detection. Coupled with a density-based outlier detection algorithm, SEG achieves competitive results on three real task-oriented dialogue datasets in two languages for unknown intent detection. On top of that, we propose to integrate SEG as an unknown intent identifier into existing generalized zero-shot intent classification models to improve their performance. A case study on a state-of-the-art method, ReCapsNet, shows that SEG can push the classification performance to a significantly higher level.

1 Introduction

Understanding user intent is crucial for developing conversational and dialogue systems. It is essential to accurately identify the intent behind a user utterance to better guide downstream decisions and policies. With the advent of conversational AI, dialogue systems are becoming central tools in many applications such as mobile apps, companion bots, virtual assistants and so on. Since user interests may change frequently over time, the AI agents may continuously see unknown (new) user intents. Manual annotation can hardly catch up with such rapid development, which motivates the problem of unknown intent detection that has recently attracted increasing interest from both academia and industry.

While there have been some pioneering works studying the open-world classification problem in natural language processing (Fei and Liu, 2016; Shu et al., 2017), very few methods are designed for unknown intent detection. To our knowledge, the first work is by Lin and Xu (2019), in which the authors use large margin cosine loss (LMCL) to learn deep discriminative features and then feed them to a density-based outlier detection algorithm to identify unknown intents. Although this method performs well on some benchmark datasets, it has two limitations. (1) In training, LMCL ignores the prior knowledge of class labels, while it has been shown that label correlations captured in the embedding space can improve prediction performance, especially in the zero-shot learning scenarios (Palatucci et al., 2009; Ma et al., 2016). (2) LMCL computes the cosine distance between embeddings in the feature space and trains with a softmax cross-entropy loss, making the embedding distribution of each class long and narrow (Wan et al., 2018), which may be less suitable for applying density-based anomaly detection algorithms to detect unknown intents.

In this paper, we aim to address these limitations and propose a novel semantic-enhanced Gaussian mixture model (SEG) for unknown intent detection. In contrast to the softmax function, the Gaussian mixture model enforces embeddings to form ball-like dense clusters in the feature space, which may be more desirable for outlier detection, especially when using density-based outlier detection algorithms. Furthermore, we propose to inject the semantic information of class labels into the Gaussian mixture distribution by assigning the embeddings of class labels or descriptions to be the means of the Gaussians. This enables SEG to learn

[*] Equal contribution.
[†] Corresponding author.

more class-concentrated embeddings that can benefit downstream outlier detection. We further use a large margin loss to make SEG learn more discriminative features and employ a density-based outlier detection algorithm LOF (Breunig et al., 2000) to detect unknown intents.

Identifying unknown intents is not enough for some application scenarios where it is important to know what exactly the new intents are, e.g., zero-shot intent classification. Current generalized zero-shot intent classification methods (Chen et al., 2016; Kumar et al., 2017; Xia et al., 2018; Liu et al., 2019) attempt to classify test instances directly by making predictions in the pool of all the seen and unseen intents. However, their prediction performances are quite low, and they are still far from practical use. In this work, we propose to integrate SEG as an unknown intent identifier into the generalized zero-shot intent classification pipeline. The basic idea is that correctly identifying if the intent of an utterance is known or unknown will make the subsequent intent classification task much easier. We conduct a case study on a state-of-the-art zero-shot intent classification method ReCapsNet (Liu et al., 2019). The results show that incorporating SEG successfully improves the performance of ReCapsNet by a large margin. It even pushes the performance to a practical level on the SNIPS dataset (Coucke et al., 2018).

The main contributions of this paper are summarized as follows.

- We propose a semantic-enhanced Gaussian mixture model (SEG) for unknown intent detection by incorporating class semantic information into a Gaussian mixture distribution.
- We explore to improve existing generalized zero-shot intent classification systems with an unknown intent identifier. To the best of our knowledge, this is the first attempt to apply unknown intent detection in this task.
- We conduct extensive experiments on three real-world datasets to validate the effectiveness of the proposed SEG model for unknown intent detection and its application in generalized zero-shot intent classification.

The rest of the paper is organized as follows. In Section 2, we review related works on intent classification and open-world classification. In Section 3, we discuss the proposed SEG model in details. In Section 4, we present experimental results on unknown intent detection. In Section 5, we apply SEG to improve generalized zero-shot intent classification and conduct a case study. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Intent Classification

User intent classification is an important component of dialogue systems. Great effort has been made to understand user intent across various domains, ranging from search engine questions (Hu et al., 2009) to medical queries (Zhang et al., 2016). Deep learning models including convolutional neural networks (CNN) (Xu and Sarikaya, 2013) and attention-based recurrent neural networks (RNN) (Ravuri and Stolcke, 2015; Liu and Lane, 2016) are commonly used for intent classification. CNN based methods build sentence embeddings by aggregating embeddings of adjacent words, while RNN based methods extract sentence embeddings via encoding word embeddings sequentially. Both types of methods have shown promising results in practice (Yin et al., 2017).

Traditional intent classification methods require considerable amount of labeled data for each class to train a discriminative classifier, while zero-shot intent classification (Sappadla et al., 2016; Zhang et al., 2019) addresses the problem that not all intent categories are seen during the training phase, which is an important task in natural language understanding as novel intents may continuously emerge in dialogue systems (Liu and Lane, 2016; Nam et al., 2016; Xu and Sarikaya, 2013). Zero-shot intent classification aims to generalize knowledge and concepts learned from seen intents to recognize unseen intents. Early methods (Ferreira et al., 2015a,b; Yazdani and Henderson, 2015) explore the relationship between seen and unseen intents by introducing external resources such as manually defined attributes or label ontologies, but they are usually expensive to obtain. To deal with this, some methods (Chen et al., 2016; Kumar et al., 2017) map the utterances and intent labels to an embedding space and then model their relations in the space. Recently, IntentCapsNet-ZS (Xia et al., 2018) extends capsule networks (Sabour et al., 2017) for zero-shot intent classification by transferring the prediction vectors from seen classes to unseen classes. ReCapsNet (Liu et al., 2019) shows that IntentCapsNet-ZS hardly recognizes

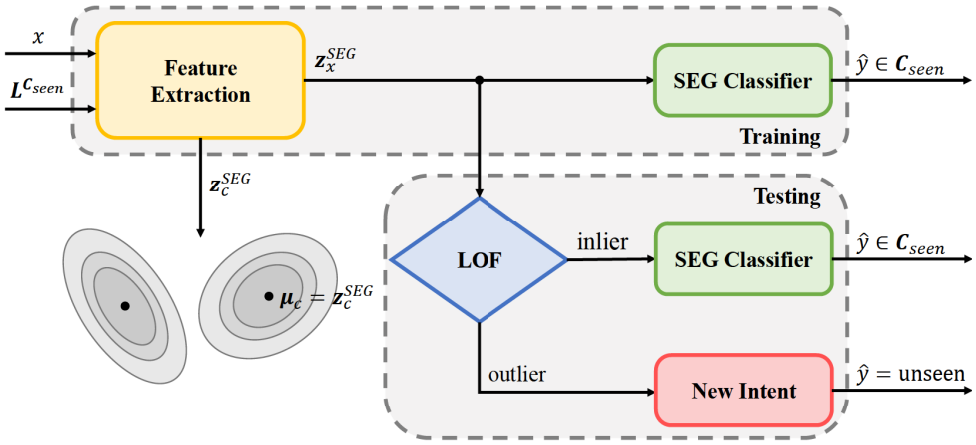

Figure 1: Illustration of the proposed framework for unknown intent classification. The backbone network is a self-attention Bi-LSTM encoder, which is trained by the proposed semantic-enhanced large margin Gaussian mixture loss (SEG classifier). In the testing phase, LOF is employed to detect outliers. The predicted outliers will be considered as unseen intent class instances, while the inliers will be classified by the SEG classifier.

utterances from unseen intents in the generalized zero-shot classification scenario, and proposes to solve this issue by transferring the transformation matrices from seen intents to unseen intents. In this paper, we use ReCapsNet as an example to show that incorporating an unknown intent identifier in the generalized zero-shot classification pipeline can significantly improve the prediction performance on unseen intents and the overall performance.

2.2 Open-world Classification

Most of existing classification methods make the closed-world assumption, that is, no new classes can appear in testing. However, the real world is open and dynamic, and in many applications, the AI agent cannot expect it sees everything in training, which makes open-world learning or classification an important problem.

There are two major approaches to tackle open-world classification. One is to use the classifier to output an additional confidence score to measure the probability that a test sample is seen or unseen. cbsSVM (Fei and Liu, 2016) proposes a center-based similarity (CBS) learning strategy and employs SVM to build 1-vs-rest CBS classifiers. MSP (Hendrycks and Gimpel, 2017) proposes to use the maximum softmax probability as the confidence score. Instead of using Softmax as the final output layer, DOC (Shu et al., 2017) builds a multi-class classifier with a 1-vs-rest final layer which contains a sigmoid function for each seen class to reduce the open space risk.

The other approach is to treat the open-world classification as an outlier detection problem by exploiting anomaly detection methods such as robust covariance estimators (Rousseeuw and Driessen, 1999), one-class SVM (Schölkopf et al., 2001), isolation forest (Liu et al., 2008) and local outlier factor (Breunig et al., 2000). Robust covariance estimators assume data follows a Gaussian mixture distribution. Based on this, it tries to fit an elliptic envelope, and outliers can be defined as points standing far enough from the fit shape. One-class SVM finds a hyperplane that circles the positive samples as the decision boundary. Isolation forest uses a binary search tree (isolated tree) to isolate samples. Due to the small number of outliers and their alienation from most samples, outliers will be isolated earlier and be closer to the root node of the isolated tree. Local outlier factor (LOF) is a density-based algorithm, which compares the density of a point and its neighbors to determine whether it is an abnormal point. Lower density means it is more likely to be identified as an abnormal point. In addition, to facilitate anomaly detection, some methods (Lin and Xu, 2019; Wan et al., 2018) use large margin loss functions to learn more discriminative feature representations.

3 Our Approach

3.1 Feature Extraction

Given an utterance $\boldsymbol{x} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_T\}$ with T words, where $\boldsymbol{w}_t \in \mathbb{R}^{d_w}$ is the embedding of the t-th word. Each word can be further encoded sequentially using a bidirectional LSTM (BiLSTM),

i.e.,

$$\begin{aligned}\overrightarrow{\boldsymbol{h}}_t &= \text{LSTM}_{fw}(\boldsymbol{w}_t, \overrightarrow{\boldsymbol{h}}_{t-1}),\\ \overleftarrow{\boldsymbol{h}}_t &= \text{LSTM}_{bw}(\boldsymbol{w}_t, \overleftarrow{\boldsymbol{h}}_{t+1}),\end{aligned} \tag{1}$$

where $\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t \in \mathbb{R}^{d_h}$ are the hidden states of the word $\boldsymbol{w}_t$ by forward LSTM_{fw} and backward LSTM_{bw} respectively. The word $\boldsymbol{w}_t$ is encoded as the entire hidden state, which is represented by concatenating $\overrightarrow{\boldsymbol{h}}_t$ and $\overleftarrow{\boldsymbol{h}}_t$, i.e. $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t; \overleftarrow{\boldsymbol{h}}_t]$, and the hidden state matrix of the utterance can be represented as $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T] \in \mathbb{R}^{2d_h \times T}$. Furthermore, we use the self-attention mechanism to obtain the sentence embedding. Specifically,

$$\begin{aligned}\boldsymbol{a} &= \text{softmax}\left(\boldsymbol{W}_{s2}\text{tanh}(\boldsymbol{W}_{s1}\boldsymbol{H})\right),\\ \boldsymbol{z} &= \boldsymbol{W}\boldsymbol{H}\boldsymbol{a},\end{aligned} \tag{2}$$

where $\boldsymbol{a} \in \mathbb{R}^T$ is the self-attention weight vector, $\boldsymbol{W}_{s1} \in \mathbb{R}^{d_a \times 2d_h}$ and $\boldsymbol{W}_{s2} \in \mathbb{R}^{1 \times D_a}$ are trainable parameters, $\boldsymbol{W} \in \mathbb{R}^{d_z \times 2d_h}$ is also trainable feed-forward weight parameter, and $\boldsymbol{z} \in \mathbb{R}^{d_z}$ is the final representation of the utterance $\boldsymbol{x}$.

3.2 Semantic-Enhanced Large Margin Gaussian Mixture Loss

The softmax cross-entropy loss is widely used in many machine learning problems. However, the embedding distribution of each class learned by the softmax cross-entropy loss tends to be long, narrow, and radiating from the center, with different classes distributed next to each other closely (Wan et al., 2018). Such embedding distribution may not be ideal for detecting new intent classes, as there might not be much space for new classes. Nevertheless, the Gaussian mixture loss can enforce each class to gather into a dense and small cluster, which may be more desirable for detecting new intents. Here, we design a semantic-enhanced large margin Gaussian mixture loss for embedding learning.

Large-Margin Cross-Entropy Loss Given a K-way classification task, we assume the extracted feature vector (embedding) $\boldsymbol{z}$ of the training samples follows a Gaussian mixture distribution, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance of class k in the embedding space respectively and $p(k)$ is the prior probability of class k. The probability density function of $\boldsymbol{z}$ is given by

$$p(\boldsymbol{z}) = \sum_k \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(k), \tag{3}$$

where $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian distribution.

For the embedding $\boldsymbol{z}_i$ of any training sample $\boldsymbol{x}_i$, the posterior probability that $\boldsymbol{z}_i$ belongs to its class y_i can be expressed as

$$p(y_i|\boldsymbol{z}_i) = \frac{\mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma}_{y_i})p(y_i)}{\sum_k \mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(k)}. \tag{4}$$

The cross-entropy loss of $\boldsymbol{z}_i$ between the true class label y_i and the inference $p(y_i|\boldsymbol{z}_i)$ can then be computed as:

$$\mathcal{L}_{ce,i} = -\log p(y_i|\boldsymbol{z}_i), \tag{5}$$

and the total loss of N training samples is

$$\mathcal{L}_{ce} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{ce,i}. \tag{6}$$

Let d_k be the Mahalanobis distance between $\boldsymbol{z}_i$ and $\boldsymbol{\mu}_k$, i.e.,

$$d_k = (\boldsymbol{z}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu}_k)/2. \tag{7}$$

Then $\mathcal{L}_{ce,i}$ can be expressed as

$$\mathcal{L}_{ce,i} = -\log \frac{p(y_i)|\boldsymbol{\Sigma}_{y_i}|^{-\frac{1}{2}} e^{-d_{y_i}}}{\sum_k p(k)|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} e^{-d_k}}. \tag{8}$$

Consider a simplified case where $p(k)$ and $\boldsymbol{\Sigma}_k$ are identical for all classes. In this case, the model will give a correct prediction of $\boldsymbol{z}_i$ if the distance of $\boldsymbol{z}_i$ to its class mean $\boldsymbol{\mu}_{y_i}$ is less than or equal to its distance to any other class mean.

In general, large margin loss helps to improve classification performance. Here, we also introduce a classification margin $m \in [1, +\infty)$ into the cross-entropy loss, which then becomes:

$$\begin{aligned}\mathcal{L}_{ce}^{m} &= \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{ce,i}^{m},\\ \mathcal{L}_{ce,i}^{m} &= -\log \frac{p(y_i)|\boldsymbol{\Sigma}_{y_i}|^{-\frac{1}{2}} e^{-md_{y_i}}}{\sum_k p(k)|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} e^{-d_k}}.\end{aligned} \tag{9}$$

With the large margin loss, $\boldsymbol{z}_i$ is correctly classified only when its distance to class mean $\boldsymbol{\mu}_{y_i}$ is significantly less than (no more than $\frac{1}{m}$ of) its distance to any other class mean.

Semantic Enhancement via Class Description This is one of the key features of our proposed method. We inject the semantic information of each class into the Gaussian mixture model by assigning

the embedding learned from the text description d_k of class k to be the class centroid $\boldsymbol{\mu}_k$. The text description d_k can either be a single-word class name or a sentence or paragraph that describes the class. That is,

$$\boldsymbol{\mu}_k = \text{feature\_extract}(d_k), \tag{10}$$

where feature_extract$(\cdot)$ indicates the feature extraction module in Section 3.1.

Generation Loss In addition to the cross-entropy loss, we want to maximize the observed likelihood of the embeddings with the Gaussian mixture distribution. Specifically, we minimize the following negative logarithm likelihood,

$$\begin{aligned}\mathcal{L}_g =& -\sum_{i=1}^{N} \log \mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma}_{y_i}) \\ =& \frac{1}{2}\sum_{i=1}^{N} (\boldsymbol{z}_i - \boldsymbol{\mu}_{y_i})^\top \boldsymbol{\Sigma}_{y_i}^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_{y_i}) \\ &+ \text{const},\end{aligned} \tag{11}$$

where const means a constant number. As shown in Eq. (11), the generation loss $\mathcal{L}_g$ encourages the embedding $\boldsymbol{z}_i$ to be close to its class centroid $\boldsymbol{\mu}_{y_i}$, which facilitaes learning a more class-concentrated embedding distribution that may benefit the downstream outlier detection task.

By combining the cross-entropy loss and the generation loss, the total objective function is:

$$\mathcal{L} = \mathcal{L}_{ce}^{m} + \lambda \mathcal{L}_g, \tag{12}$$

where λ is a trade-off parameter.

3.3 Outlier Detection

By the above feature learning procedure, each utterance $\boldsymbol{x}$ can be encoded as an embedding $\boldsymbol{z}$. Then, the embedding $\boldsymbol{z}$ is fed to a well-known outlier detection algorithm LOF (Breunig et al., 2000) to detect new or unknown intents (outliers). LOF is an unsupervised density-based anomaly detection method based on the following intuition. By comparing the local density of an object to those of its neighbors, it can identify regions of similar density. The objects with substantially lower density than their neighbors' are considered to be outliers.

LOF defines the local outlier factor of an object $\boldsymbol{z}$ as

$$\text{LOF}_k(\boldsymbol{z}) = \frac{1}{|N_k(\boldsymbol{z})|} \sum\nolimits_{\boldsymbol{o} \in N_k(\boldsymbol{z})} \frac{\text{lrd}_k(\boldsymbol{o})}{\text{lrd}_k(\boldsymbol{z})}, \tag{13}$$

Dataset	SNIPS	ATIS	SMP-2018
Vocab Size	11642	938	3189
Avg. Length	9.05	11.37	4.87
# of Samples	13802	6371	2460
# of Classes	7	18	30

Table 1: Dataset statistics.

where $N_k(\boldsymbol{z})$ denotes the set of k-nearest neighbors of $\boldsymbol{z}$, and "lrd" denotes the *local reachability density* which measures the local density around an object. The local reachability density is defined as the inverse of the average reachability distance between $\boldsymbol{z}$ and its neighbors, i.e.,

$$\text{lrd}_k(\boldsymbol{z}) = \frac{|N_k(\boldsymbol{z})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{z})} \text{reach-dist}_k(\boldsymbol{z}, \boldsymbol{o})}. \tag{14}$$

Here, the reachability distance reach-dist$_k(\boldsymbol{z}, \boldsymbol{o})$ is defined as

$$\text{reach-dist}_k(\boldsymbol{z}, \boldsymbol{o}) = \max\left\{\text{k-dist}(\boldsymbol{o}), d(\boldsymbol{z}, \boldsymbol{o})\right\}, \tag{15}$$

where k-dist$(\boldsymbol{o})$ denotes the distance of the object $\boldsymbol{o}$ to its k-th nearest neighbor, and $d(\boldsymbol{z}, \boldsymbol{o})$ is the distance between $\boldsymbol{z}$ and $\boldsymbol{o}$.

If the LOF factor of an utterance is much larger than 1, it has substantially lower local density than its neighbors', which means the utterance embedding is relatively distant from its neighbors. Hence, it can be inferred the utterance is likely to belong to an unknown intent class.

3.4 Overall Procedures

Figure 1 illustrates the overall training and testing procedures of the proposed framework for unknown intent detection. The backbone network is a self-attention Bi-LSTM encoder. In the training phase, the encoder is trained by minimizing the semantic-enhanced large margin Gaussian mixture loss (SEG classifier) as in Eq. (12) on the training samples (seen intent class instances). In the testing phase, user utterances may come from both seen and unseen intent classes. Given an utterance, we first obtain its feature representation $\boldsymbol{z}$ with the trained encoder, then we use LOF to decide whether $\boldsymbol{z}$ is an outlier or not. If $\boldsymbol{z}$ is an outlier, we take it as an instance of some new intent class. Otherwise, we classify $\boldsymbol{z}$ to one of the seen intent classes using the SEG classifier.

4 Experiments

In this section, we present experimental results on unknown intent detection. Formally, we train an

Dataset	SNIPS			ATIS			SMP-2018		
% of known intents	25%	50%	75%	25%	50%	75%	25%	50%	75%
MSP	**0.5543**	0.8060	**0.8585**	**0.6848**	0.5158	0.3853	0.6132	0.7089	0.7716
DOC	0.5462	0.7962	0.8564	**0.7007**	0.5073	0.3659	0.6095	0.7197	0.7642
Softmax	0.5508	0.8036	0.8393	0.6597	0.6310	0.5732	0.5818	0.6860	0.7351
LMCL	0.5489	0.8041	0.8458	0.6763	**0.6778**	**0.6110**	0.6059	0.7094	0.7580
SEG/o	0.5440	**0.8067**	0.8474	0.6768	0.6699	0.5918	**0.6734**	**0.7676**	**0.8128**
SEG	**0.5599**	**0.8193**	**0.8612**	0.6410	**0.6700**	**0.6466**	**0.6966**	**0.7895**	**0.8205**

Table 2: Macro F1-score of unknown intent detection with different proportion of seen classes. The top 2 results for each metric are marked in bold.

unknown intent detection system with training data $D^{tr} = (X^{tr}, Y^{tr})$, where $Y^{tr} \in \{l_1, \cdots, l_K\} = C_{\text{seen}}$ (the set of seen intent classes). For test utterances of seen intents, the unknown intent detection system aims to assign correct intent labels to them. For test utterances of unseen intents, the system is expected to identify them as outliers.

4.1 Datasets and Baselines

We evaluate our method SEG for unknown intent detection on 3 real task-oriented dialogue datasets: SNIPS (Coucke et al., 2018), ATIS (Hemphill et al., 1990) and SMP-2018 (Zhang et al., 2017). SNIPS is an open-source single-turn English corpus, which contains 7 types of user intents across different domains. ATIS is also an English dataset, which contains 18 types of user intent in the airline travel domain. SMP-2018 is a Chinese dialogue corpus for user intent recognition, which contains 30 different types of user intents. The statistics of the datasets are summarized in Table 1.

We compare SEG with the following unknown intent detection methods.

- **Maximum Softmax Probability (MSP)** (Hendrycks and Gimpel, 2017) considers the maximum softmax probability of a sample as the confidence score to measure the probability that it belongs to a seen intent. The smaller the confidence score is, the more likely it belongs to an unknown intent.

- **DOC** (Shu et al., 2017) builds m 1-vs-rest sigmoid classifiers for m seen classes respectively. The maximum probability is considered as the confidence of whether the sample belongs to the seen intent.

- **Softmax**. It can be considered as an ablation study of our method SEG, which uses softmax instead of Gaussian mixture distribution to learn discriminative features.

- **LMCL** (Lin and Xu, 2019) uses large margin cosine loss instead of Gaussian mixture distribution to learn discriminative embeddings.

- **SEG/o**. A variant of our method SEG. It does not inject the class semantic information into the Gaussian mixture model.

4.2 Experimental Setup

We follow the setting in LCML (Lin and Xu, 2019) for unknown intent detection. Considering that some datasets may be unbalanced, we randomly select seen intents by a weighted random sampling over the entire intent set. The rest of the intents are regarded as unknown. We randomly select 30% samples of each intent to form the test set. The rest of each seen intent is added to the training set. We also follow LMCL to use macro f1-score as the evaluation metric, which makes sense because the ATIS dataset is extremely unbalanced.

For SNIPS, ATIS and SMP-2018, we use 300-dim embeddings pre-trained on Fasttext, Glove, and Chinese-Word-Vectors respectively. For BiLSTM, we set the number of layers as 2 and the output dimension as 128. In the self-attention layer, we set the attention dimension d_a=10. After the self-attention layer, we project the feature vector to a d_z-dimension vector via a linear layer. We set d_z=12 for SNIPS and SMP-2018, and d_z=4 for ATIS. We report the average results over 10 runs. For the loss function, we set the margin $m = 1$ and the trade-off parameter $\lambda = 0.5$.

For MSP, we set the threshold as 0.5 following Lin and Xu (2019). For DOC, we set the threshold as 0.5 as used in the original paper. During training of MSP and DOC, we clip the gradient norm to avoid gradient exploding. For LMCL, we follow

the original paper to set the scaling factor $s = 30$ and the cosine margin $m = 0.35$. Softmax, LMCL, SEG/o and SEG all use LOF as the outlier detector, and we use the same set of parameters for LOF.

4.3 Result Analysis

From Table 2, it can be seen that our method SEG outperforms the baselines in most cases. Especially, on the most challenging dataset SMP-2018, SEG and SEG/o outperfom others by a large margin, demonstrating its high effectiveness. Moreover, we can make the following observations:

(1) SEG consistently outperforms SEG/o in most cases, which proves the effectiveness of the proposed semantic enhancement mechanism.

(2) SEG/o generally has higher scores than Softmax and LMCL, especially on the more complex dataset SMP-2018, where significant gaps can be observed. The results indicate the advantage of Gaussian mixture model over Softmax and the variant LMCL for learning class-concentrated embeddings, which are more suitable to be coupled with the outlier detector LOF.

(3) All the methods work well on SNIPS, which is a simple dataset. MSP and DOC outperform other methods on ATIS with only 25% seen classes. However, as the proportion of seen class increases, we can see a significant decline in their performance. This is because ATIS is severely imbalanced where one intent accounts for 96% of the entire data. When there are many seen classes, DOC and MSP cannot learn an effective supervised classifier due to the dominance of one class.

5 Application in Generalized Zero-shot Intent Classification

In this section, we apply our method SEG to an extended application of unknown intent classification – zero-shot intent classification. It aims to discriminate unseen intents, which is beyond only detecting their existence. Specifically, given the training data $D^{tr} = (X^{tr}, Y^{tr})$ where $Y^{tr} \in C_{\text{seen}}$, a zero-shot classification system is trained to predict the label $\hat{y}^{te}$ of any test sample which may belong to an unseen class, using the knowledge transferred from the seen data. There are two common settings for zero-shot learning, generalized zero-shot classification, where $\hat{y}^{te} \in \{C_{\text{seen}}, C_{\text{unseen}}\}$, and standard zero-shot classification, where $\hat{y}^{te} \in C_{\text{unseen}}$. Here, C_{unseen} is the set of unseen intent classes.

Previous attempts try to tackle the challenge of

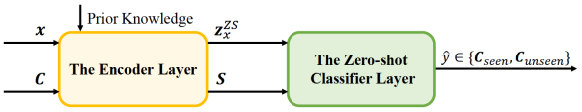

Figure 2: A typical generalized zero-shot intent classification pipeline.

zero-shot intent classification from three directions. (1) What prior knowledge is more supportive, such as morphology (character-level embedding), class descriptions, and knowledge-based entity attributes (Ferreira et al., 2015a,b; Chen et al., 2016; Kumar et al., 2017). (2) How to better utilize these prior knowledge to extract more informative semantic representations, such as data augmentation and hierarchical representations learned by capsule networks (Xia et al., 2018). (3) With the extracted semantic features, how to design a better zero-shot learning strategy, such as reconstructing weight matrix for unseen intents through relation learning (Liu et al., 2019).

In this work, we improve generalized zero-shot intent classification by integrating the proposed SEG model as a binary unknown intent identifier into the original pipeline. We explore multiple ways of integration and conduct a case study based on a state-of-the-art method ReCapsNet (Liu et al., 2019).

5.1 Integrating Unknown Intent Identifier

As shown in Figure 2, a typical generalized zero-shot classification framework can be abstracted into two layers, the encoder layer and the zero-shot classifier layer. In the encoder layer, a user utterance x in the text format needs to be first mapped to the semantic representation $\boldsymbol{z}_x^{ZS}$. In addition, it is common to encode class information as $\boldsymbol{S}$ for better semantic learning or knowledge transfer. In order to learn better semantic representation, prior knowledge is usually incorporated at this stage. Then, the learned representation will be fed to the zero-shot classifier layer. Various zero-shot classification strategies have been proposed to transfer knowledge to new categories. Finally, the system outputs the prediction $\hat{y}^{te} \in \{C_{\text{seen}}, C_{\text{unseen}}\}$ for the utterance x.

We integrate SEG into the pipeline between the encoder layer and the classifier layer as shown in Figure 3. With the semantic feature $\boldsymbol{z}_x$, we predict if the utterance x is an outlier via:

$$p(g|\boldsymbol{z}_x), g \in \{\text{“seen”}, \text{“unseen”}\}. \quad (16)$$

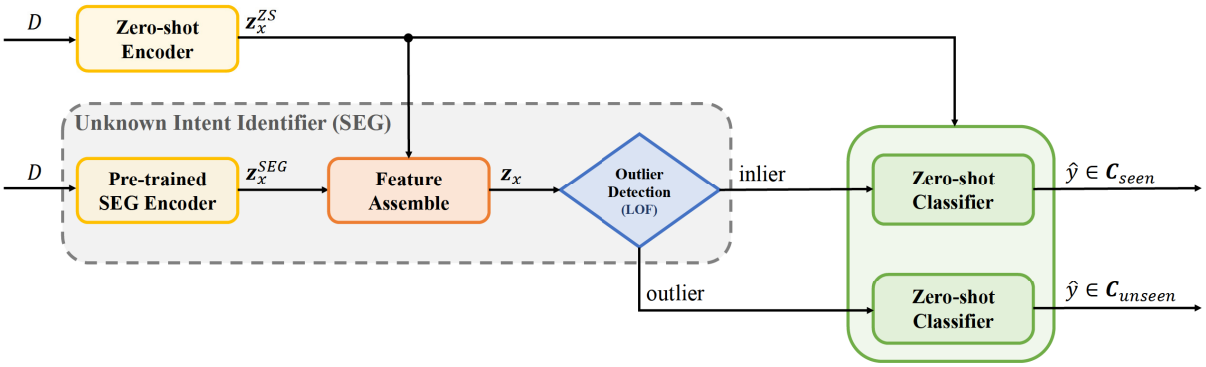

Figure 3: Integration of the new intent identifier (SEG) into the generalized zero-shot intent classification pipeline.

For the case g = "seen", the intent of the utterance is considered to be a seen one. We then predict the intent by $p(y|\boldsymbol{z}_x, y \in C_{\text{seen}}, X^{tr}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ denotes the parameters of the original framework. Otherwise, the intent of the utterance is considered to be unseen, and we predict it via $p(y|\boldsymbol{z}_x, y \in C_{\text{unseen}}, X^{tr}, \boldsymbol{\theta})$.

Feature Assemble We adopt two ways "Separate" and "Combine" to assemble features for the following outlier detection task.

- **Separate (Sep).** We directly feed the output of the pre-trained SEG encoder $\boldsymbol{z}_x^{SEG}$ to LOF for outlier detection, i.e.,

$$\boldsymbol{z}_x = \boldsymbol{z}_x^{SEG}. \tag{17}$$

- **Combine.** To take advantage of the original model, we first obtain the original semantic feature representation $\boldsymbol{z}_x^{ZS}$ and define a transform function f. Then, $f(\boldsymbol{z}_x^{ZS})$ is concatenated with the pre-trained features by SEG, $\boldsymbol{z}_x^{SEG}$, to make a combined feature representation:

$$\boldsymbol{z}_x = [\boldsymbol{z}_x^{SEG} || f(\boldsymbol{z}_x^{ZS})]. \tag{18}$$

5.2 A Case Study on ReCapsNet

ReCapsNet Recently, ReCapsNet-ZS (Liu et al., 2019) demonstrates state-of-the-art performance in generalized zero-shot intent classification. In this section, we conduct a case study on integrating the new intent identifier into ReCapsNet.

The framework of ReCapsNet is illustrated in Figure 4. In the encoder layer, each utterance x is encoded with R semantic capsules $[\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_R]$ as the representations in R different semantic spaces. In addition, the training set $\boldsymbol{D}^{tr}$ and class labels $\boldsymbol{L}$ are encoded as $\boldsymbol{S}^{tr}$ and

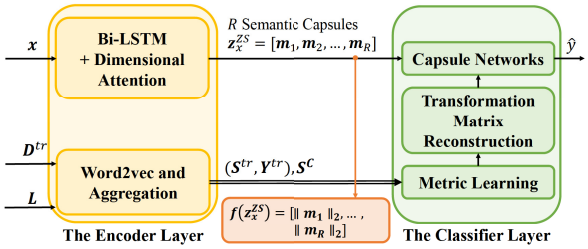

Figure 4: The framework of ReCapsNet.

$\boldsymbol{S}^C$, respectively. In the zero-shot classifier layer, $\boldsymbol{z}_{\boldsymbol{x}}^{ZS}$ is fed to a capsule network to make prediction. Each seen class k has R transformation matrices $\{\boldsymbol{W}_{kr}\}_{r=1}^{R}$. In the testing phase, ReCapsNet reconstructs the r-th transformation matrix for each unseen class l as $\boldsymbol{W}_{lr} = \sum_k q_{lk} \boldsymbol{W}_{kr}$, where q_{lk} is the relation between unseen class l and seen class k learned from $(\boldsymbol{S}^{tr}, \boldsymbol{Y}^{tr})$ and $\boldsymbol{S}^C$ by metric learning.

For the variant "Combine", to exploit the property that each utterance is variously represented in different semantic spaces as discussed in Liu et al. (2019), we define the semantic feature representation of ReCapsNet as

$$f(\boldsymbol{z}_x^{ZS}) = [\|\boldsymbol{m}_1\|_2, \|\boldsymbol{m}_2\|_2, \cdots, \|\boldsymbol{m}_R\|_2]. \tag{19}$$

Experimental Setup We integrate SEG into the ReCapsNet pipeline with both "Sep" and "Combine" variants and test the performance of generalized zero-shot classification.

Following the settings of generalized zero-shot classification in Liu et al. (2019), we test our methods on two datasets SNIPS (Coucke et al., 2018) and SMP-2018 (Zhang et al., 2017) and report the micro-averaged recall (accuracy) and F1 scores. The baselines include DeVISE (Frome et al., 2013), CMT (Socher et al., 2013), CDSSM (Chen et al., 2016), Zero-shot DNN (Kumar et al., 2017), Intent-

Method	SNIPS						SMP-2018					
	Seen		Unseen		Overall		Seen		Unseen		Overall	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DeViSE	0.9481	0.6536	0.0211	0.0398	0.4215	0.3049	0.8040	0.6740	0.0270	0.0310	0.5030	0.4250
CMT	**0.9755**	0.6648	0.0397	0.0704	0.4438	0.3271	**0.8314**	0.7221	0.0798	0.1069	0.5398	0.4834
CDSSM	0.9549	0.7033	0.0111	0.0218	0.4234	0.3194	0.6653	0.5540	0.1436	0.1200	0.4864	0.4052
Zero-shot DNN	0.9432	0.6679	0.0682	0.1041	0.4488	0.3493	0.7323	0.6116	0.0590	0.0869	0.5013	0.4316
IntentCapsNet	**0.9741**	0.6517	0.0000	0.0000	0.4200	0.2810	**0.8850**	0.7281	0.0000	0.0000	0.5375	0.4423
ReCapsNet	0.9511	0.6777	0.0994	0.1594	0.4705	0.3826	0.8107	**0.7417**	0.1959	0.1727	0.5692	0.5182
SEG (Sep / o)	0.9308	0.7501	0.3523	0.4514	0.6014	0.5800	0.7066	**0.7391**	0.3848	0.3038	0.5802	0.5681
SEG (Combine / o)	0.9217	0.7924	0.4642	0.5321	0.6612	0.6441	0.7054	0.7326	0.3888	0.3116	0.5811	0.5672
SEG (Sep / w)	0.7898	**0.8335**	**0.6728**	**0.6420**	**0.7232**	**0.7245**	0.6624	0.7243	**0.4779**	**0.3627**	**0.5899**	**0.5823**
SEG (Combine / w)	0.8644	**0.8658**	**0.6961**	**0.6931**	**0.7685**	**0.7674**	0.6821	0.7359	**0.4848**	**0.3806**	**0.6046**	**0.5963**

Table 3: Results of generalized zero-shot intent classification equipped with our new intent identifier SEG. "Seen", "Unseen" and "Overall" respectively denote the prediction performance on the utterances from seen intents, unseen intents, and both seen and unseen intents. The suffixes "/w" and "/o" stand for with and without semantic enhancement, respectively. The top 2 results for each metric are marked in bold.

CapsNet (Xia et al., 2018), and ReCapsNet (Liu et al., 2019). The average results over 10 runs of our methods and ReCapsNet are reported in Table 3, where the results of other baselines are taken from Liu et al. (2019).

We use the same setting and hyper-parameters as in ReCapsNet (Liu et al., 2019). We set d_z=4 for SNIPS and d_z=12 for SMP-2018. The rest of the parameters of SEG are the same as those used in Section 4.2. In addition, we also conduct an ablation study to demonstrate the effectiveness of the proposed semantic enhancement mechanism by testing two variants of our integration ("Sep / o" and "Combine / o") without using it.

Result Analysis From the results in Table 3, we can make the following observations:

(1) All variants of our integration achieve a significant boost in the overall accuracy and F1 scores on the two datasets, especially on SNIPS, where the performance increase is huge. Each variant leads to a qualitative leap in the performance on unseen intents. The prediction accuracy (micro-averaged recall) on seen intents may be reduced compared to ReCapsNet and other baselines, since some utterances of seen intents are classified to unseen intents. However, the F1 score on seen intents increases significantly, indicating that it has much higher precision than that of the baselines.

(2) The variants of our integration with semantic enhancement significantly outperform those without using it on predicting unseen intents by very large margins. Although their accuracy scores on seen intents are lower, their overall accuracy and F1 scores are consistently better, which confirms the effectiveness of semantic enhancement.

(3) It can be seen that the "Combine" variants generally perform much better than the "Sep" variants, especially the one with semantic enhancement ("Combine / w"), which performs outstandingly. It surpasses the performance of "Sep / w" in every metric, demonstrating the usefulness of the simple feature assemble strategy of concatenating the feature representations of ReCapsNet and SEG.

6 Conclusion

In this paper, we have proposed SEG, a semantic-enhanced Gaussian mixture model coupled with a LOF outlier detector, for unknown (new) intent detection. We empirically verified the effectiveness of SEG for unknown intent detection on real dialogue datasets in English and Chinese. Furthermore, we successfully applied SEG to improve generalized zero-shot intent classification and achieved remarkable performance gain over a most recent competitive method ReCapsNet. In future work, we plan to conduct more empirical studies on SEG and further improve its performance on new intent identification. We also plan to conduct more case studies in applying SEG to boost the performance of current zero-shot intent classification methods.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This research was supported by the grant HK ITF UIM/377.

References

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. Lof: identifying density-based local outliers. In *ACM SIGMOD RECORD*, volume 29, pages 93–104.

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

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.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 506–514.

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

Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefèvre. 2015b. Zero-shot semantic parser for spoken language understanding. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1403–1407.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2121–2129.

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, USA, June 24-27, 1990*. Morgan Kaufmann.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*.

Jian Hu, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. 2009. Understanding user's query intent with wikipedia. In *International Conference on World Wide Web (WWW)*, pages 471–480.

Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Björn Hoffmeister. 2017. Zero-shot learning across heterogeneous overlapping domains. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2914–2918.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5491–5496. Association for Computational Linguistics.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 685–689.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *IEEE International Conference on Data Mining (ICDM)*, pages 413–422.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4798–4808.

Yukun Ma, Erik Cambria, and Sa Gao. 2016. Label embedding for zero-shot fine-grained named entity typing. In *International Conference on Computational Linguistics (COLING)*, pages 171–180.

Jinseok Nam, Eneldo Loza Menc'ia, and Johannes Fürnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1948–1954.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1410–1418.

Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 135–139.

Peter J Rousseeuw and Katrien Van Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3859–3869.

Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Menc'ia, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *European Symposium on Artificial Neural Networks (ESANN)*.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: deep open classification of text documents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2911–2916. Association for Computational Linguistics.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 935–943.

Weitao Wan, Yuanyi Zhong, Tianpeng Li, and Jiansheng Chen. 2018. Rethinking feature distribution for loss functions in image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9117–9126.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3090–3099.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU Workshop)*, pages 78–83.

Majid Yazdani and James Henderson. 2015. A model of zero-shot learning of spoken language understanding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 244–249.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv*, abs/1702.01923.

Chenwei Zhang, Wei Fan, Nan Du, and Philip S. Yu. 2016. Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach. In *International Conference on World Wide Web (WWW)*, pages 1373–1384.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1031–1040.

Weinan Zhang, Zhigang Chen, Wanxiang Che, Guoping Hu, and Ting Liu. 2017. The first evaluation of chinese human-computer dialogue technology. *arXiv*, abs/1709.10217.