Intent Detection and Zero-shot Intent Classification for Chatbots

Sobha Lalitha Devi and Pattabhi RK Rao
AU-KBC Research Centre,
MIT Campus of Anna University, Chennai
sobha@au-kbc.org, pattabhi@au-kbc.org

Abstract

In this paper we give in detail how seen and unseen intent is detected and classified. User intent detection has a critical role in dialogue systems. While analysing the intents it has been found that intents are diversely expressed and new variety of intents emerge continuously. Here we propose a capsule-based approach that classifies the intent and a zero-shot learning to identify the unseen intent. There are recently proposed methods on zero-shot classification which are implemented differently from ours. We have also developed an annotated corpus of free conversations in Tamil, the language we have used for intent classification and for our chatbot. Our proposed method on intent classification performs well.

1 Introduction

With the emergence of conversational AI, chatbots or dialogue systems are becoming central tools in many applications such as virtual assistants helping the citizens in getting authentic information from governments, academia, health sector and industry and so on. Since user interests may change frequently over time, the AI agents may continuously see unknown (new) user intents. It is necessary and essential to understand the intent and accurately identify the intent behind the user utterance. This will help in generating the correct response for the user intent. Data created manually by manual annotation cannot catch up with the requirement and algorithms which use annotated data cannot give accurate understanding of the intent by the user. This motivates the problem of detection of intent and its classification and has recently attracted increasing interest from both academia and industry. Here we define Intent as user's intention in an utterance which is the Turn Constructional Unit (TCU), the fundamental unit, in an utterance.

In this paper, we aim to address the issue of detection of intent, seen and unseen intent and its classification. Here we tackle this issue using zero-shot intent detection and classification problem with a capsule based (Hiton et.al. 2011, Sabour et.al. 2017) model, as tried by (Xia et.al. 2018). A capsule encase a vector representation of a group of neurons, and the orientation of the vector encodes properties of an object (like in intent Turn Composition Unit (TCU)), while the length of the vector reflects its probability of existence (Intent seen or unseen classification such as request). The capsule model learns a hierarchy of feature detectors via a routing-by-agreement mechanism: capsules for detecting low-level features (like in intent Turn Composition Unit (TCU)) send their outputs to high-level capsules (Intent seen or unseen classification such as request) only when there is a strong agreement of their predictions to high-level capsules.

Here we use zero-shot learning approaches for intent classification. In classification we identify the intent which is seen as well as unseen while training. There are several zero-shot learning approaches attempted to address the challenges for classifying intents whose instances are not present during training. The common method adopted to classify is to utilize some external resources (Ferreira et al., 2015a, 2015b; Yazdani and Henderson, 2015; Kumar et al., 2017; Zhang et al.,2019) like labelled ontologies or manually defined features. There is a dearth of such resources as well as developing such resources are time consuming and expensive.

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. Identifying intents is not enough for some application scenarios where it is important to know what exactly the new intents are that is the intent has to be classified such as zero-shot intent classification. To implement zero-shot intent classification more easily and intelligently, recent works use the word embedding's of intent labels, which can be easily pre-trained on text corpus. Current generalized zero-shot intent classification methods proposed by (Chen et al., 2016; Kumar et al., 2017; Xia et al., 2018; Liu et al., 2019) utilize neural networks to classify test instances directly by making predictions in the pool of all the seen and unseen intents.

The rest of the paper is organized as follows. In Section 2, we review related works on intent classification and detection. In Section 3, we give our approach along with corpus development in Tamil. The section 4 is on implementation and results followed by conclusion.

2 Related Work

There has been many studies to understand user intent from various domains, ranging from search engine questions (Hu et al., 2009) to medical queries (Zhang et al., 2016). The approaches used for intent classification include Deep learning models such as convolutional neural networks (CNN) (Xu and Sarikaya, 2013) and attention-based recurrent neural networks (RNN) (Ravuri and Stolcke, 2015; Liu and Lane, 2016). Traditional intent classification methods require considerable amount of labelled 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, but new intents could continuously emerge in dialogue systems (Liu and Lane, 2016; Nam et al., 2016; Xu and Sarikaya, 2013). Zeroshot intent classification aims to generalize knowledge and concepts learned from seen intents to recognize unseen intents. Early methods include (Ferreira et al., 2015a, 2015b; Yazdani and Henderson, 2015). Recently, IntentCapsNet-ZS (Xia et al., 2018) extends capsule networks (Sabour et al., 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 Measuring Intent Relations to tackle the issue of seen intents to unseen intents.

3 Our Approach

Here in this work we follow the two architecture approach similar to the one proposed by Xia et al, (2018). Initially we use our model, CapsNet, for intent detection and zero-shot learning for classification of unseen intent.

Here are the outline of the steps that were followed in the implementation of CapsNet intent detection:

a) Dataset preparation:
   - We have developed labelled dataset for intent detection. Each utterance is labelled with fundamental utterance unit the Turn Construction Units (TCUs) and also the complete utterance is labelled with its corresponding class of the intent.
   - Pre-process the conversation data by performing tokenization.

b) CapsNet Data Input
   - Convert the pre-processed conversation data into vector representations such that it can be fed into the neural network. Word embedding's method of Word2Vec is used to represent TCUs as dense vectors.

c) Designing the CapsNet architecture for Intent detection:
   - The original CapsNet architecture used for image recognition as used by Hinton et al 2017 is adapted to suit the intent detection task. The image-based input is replaced with the text vector representations obtained from Word2vec.
   - The number of capsule layers is derived empirically.
   - Dynamic routing-by-agreement algorithm (Sabour et al., 2017), is used for learning hierarchical relationships between TCUs and thus used for intent detection

d) Training the CapsNet:
   - The dataset is split into training and validation sets.

- CapsNet model is initialized with initial values and then use categorical cross-entropy loss function for intent detection.
- Train the model using the training dataset, monitoring the validation accuracy to avoid over-fitting.
- Backpropagation and gradient descent function to update the model's weights and thus optimize the loss function.

e) Measuring Intent Relations:

- Here we propose to learn a Mahalanobis distance metric to measure the relationship between unseen and seen intents. Specifically, given the embedding's of an unseen intent l and a seen intent k, their squared Mahalanobis distance is given by:

$$d_M(e_{ul}, e_{sk}) = (e_{ul} - e_{sk})^{>}\Omega^{-1}(e_{ul} - e_{sk}), \quad \text{---- (1)}$$

Where, Ω is a learnable covariance matrix which models the correlation between dimensions of the embedding. Note that IntentCapsNet (Xia et al., 2018) also tries to use Eq. (1) to model the relationship between unseen and seen intents, but it ignores the correlation between dimensions and simply sets Ω = σ2I (σ is a scaling hyper-parameter), which is actually a scaled squared Euclidean distance.

3.1 Zero-shot Intent Classification

Zero-shot intent classification involves using auxiliary information or semantic descriptions to associate visual features with intent labels. The zero-shot intent classification utilizes vote vectors from existing intents to build intent representations for emerging intents via a similarity metric between existing intents and emerging intents.

Suppose there are K existing intents and L emerging intents, the similarities between existing and emerging intents form a matrix Q∈R L×K. Specifically, the similarity between an emerging intent
$z_l \in Z$ and an existing intent $y_k \in Y$ is computed as:

$$q_{lk} = \frac{exp\{-d(\mathbf{e}_{z_l}, \mathbf{e}_{y_k})\}}{\sum_{k=1}^{K} exp\{-d(\mathbf{e}_{z_l}, \mathbf{e}_{y_k})\}},$$

where

$$d(\mathbf{e}_{z_l}, \mathbf{e}_{y_k}) = (\mathbf{e}_{z_l} - \mathbf{e}_{y_k})^T \Sigma^{-1} (\mathbf{e}_{z_l} - \mathbf{e}_{y_k})$$

e_{zl} , $e_{yk} \in R^{D\,1\times1}$ are intent embeddings computed by the sum of word embeddings of the intent label. Σ models the correlations among intent embedding dimensions and we use $\Sigma = \sigma^2$ I. σ is a hyper-parameter for scaling.

We feed the prediction vector n_l to Dynamic routing algorithm and derive activation vectors n_l on emerging intents as the output. The final intent representation n_l for each emerging intent is updated toward the direction where it coincides with representative votes vectors. We can easily classify the utterance of emerging intents by choosing the activation vector with the largest norm $\hat{z} = \arg\max \|n_l\|$

3.2 The Data Creation

We use the dataset developed in house. In developing the corpus we have followed the annotation convention used for developing annotated corpus of free conversations in Japanese, called "JAIST Annotated Corpus of FreeConversations"(Kiyoaki Shirai and Tomotaka Fukuoka. 2018) . Our corpus consists of dialogs of two native speakers in Tamil as participants, where they freely talk about various topics. Each utterance in the dialogs is annotated with two kinds of tags. One is a Turn construction Unit (or speech act), which is the type of utterance that represents the speaker's intention. The other is sympathy that is the interest shown by the speaker in the current topic in the conversation. The corpus consists of transcriptions of 100 free conversations between two participants. The total duration of the dialog is about 110 hours. Each utterance was transcribed by hand. Out of 100 conversations, not all were with two participants. 86 dialogs were with two people participate in the conversation. The statistics is given: Number of dialogs 86, Number of utterances 86,020, Average number of utterances per dialog 1000. This shows that each dialog is long. We had two annotators and the Inter annotator's agreement Kappa score is 92%. This Tamil conversational data is first of its kind in Indian languages.

Each utterance has the information: Speaker ID: An identification number of the speaker. Turn taking: A flag indicating whether the speaker has changed or not. TCU: A dialog act of an utterance. Sympathy tag: A tag that represents whether the speaker shows sympathy or antipathy. Nine types of TCU were formulated for the

annotation (Request, Confirmation etc.). The annotation has also three tags for sympathy.

4 Experiments & Results

We perform our experiment using the data created by us which is described in the previous section. The data is split into two, training (66 dialogues) and test (20 dialogues). The test partition is formed such that 10 dialogues have topic similarity with the training partition. The remaining 10 dialogues are completely different than the rest (completely unseen).

The embedding's needed for the intent detection models, are developed using Tamil Wikipedia content and copyright free Novels digitized content from Project Madurai (Project Madurai). These pre-trained word embedding's are used for intent classification using CapsNet and unseen internet detection using zero-shot learning. We use three fold cross-validation to choose hyper parameters. The dimension of the prediction vector D_P is 10. $D_I = D_W$ because we use the averaged word embedding's contained in the intent label as the intent embedding. An additional input dropout layer with a dropout keep rate 0.8 is applied to the intent annotated corpus of ours. In the loss function, the down-weighting coefficient $-\lambda$ is 0.5, margins m^+_k and m_k are set to 0.9 and 0.1 for all the existing intents. The iteration number *iter* used in the dynamic routing algorithm is 3. Adam optimizer is used to minimize the loss function.

4.1 Results & Evaluation

Evaluation Metrics: We use the two widely used evaluation metrics: accuracy (ACC) and micro-averaged F1 scores (F1) to evaluate the performance. We have used the baseline system using TFIDF classifier. The below table shows the results of our proposed model in comparison with our baseline model.

Table 1. Results – Intent Classifier

Method	Seen		Unseen	
	Acc	F1	Acc	F1
TFIDF Classifier (Baseline)	0.8246	0.8138	0.4453	0.4565
Zero-shot Intent Classifer (proposed system)	0.9181	0.9153	0.7823	0.7756

The results obtained are comparable with the state of the art. And it can be observed that our proposed system has obtained good results also for unseen intents.

Conclusion

In this paper we give in detail how seen and unseen intent is detected and classified. Here we base our work on CapsNet-ZS algorithm for intent detection and classification, and introduce Mahalanobis distance metric for identifying relationship between seen and unseen intent embedding's. Through the introduction of Mahalanobis distance metric for measuring intent relationships we overcome the prediction problem for unseen intents. From the results we observe that it is comparable with the state of the art systems.

References

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

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 (INTER SPEECH), pages 685–689.

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

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

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

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In ICANN, pages 44–51.

Jian Hu, Gang Wang, Frederick H. Lochovsky, JianTao Sun, and Zheng Chen. 2009.

Understanding user's query intent with wikipedia. In International Conference on World Wide Web (WWW), pages 471–480.

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

Kiyoaki Shirai and Tomotaka Fukuoka. 2018. JAIST Annotated Corpus of Free Conversation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

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 (INTER-SPEECH), pages 135–139.

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

Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

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.

Yun-Nung Chen, Dilek Z. Hakkani-T¨ur, 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.