Novel Intent Detection and Active Learning Based Classification (Student Abstract)

Ankan Mullick

Department of Computer Science and Engineering. Indian Institute of Technology Kharagpur, India
ankanm@kgpian.iitkgp.ac.in

Abstract

Novel intent class detection is an important problem in real world scenario for conversational agents for continuous interaction. Several research works have been done to detect novel intents in a mono-lingual (primarily English) texts and images. But, current systems lack an end-to-end universal framework to detect novel intents across various different languages with less human annotation effort for mis-classified and system rejected samples. This paper proposes **NIDAL** (**N**ovel **I**ntent **D**etection and **A**ctive **L**earning based classification), a semi-supervised framework to detect novel intents while reducing human annotation cost. Empirical results on various benchmark datasets demonstrate that this system outperforms the baseline methods by more than 10% margin for accuracy and macro-F1. The system achieves this while maintaining overall annotation cost to be just $\sim$ 6-10% of the unlabeled data available to the system.

Introduction

With the emerging new intents in the system, the conversational agents need to be retrained in order to identify these newer intents aka *novel classes*. Also as time progresses, the overall accuracy of identifying the known intents for the user queries should not be compromised. Currently, if a user utterance belongs to one of the known intents but the confidence score is low, these are called *rejected utterances*. It is infeasible to label all the new instances because of the sheer volume, and hence an effective mechanism to reduce human annotation cost would be required. To address these issues, in this work, a semi-supervised setting is adopted to identify known and novel intent classes. Corresponding to this problem setting, it starts with an initial labelled training data which consists of a set of known intents and a large unlabelled data that comprises of known and novel intent classes. The evaluation is performed on a predefined test set that consists of all the known and novel classes.

In the last decade, researchers have focused on out of domain class detection and explored various active learning methods in various scenarios to improve classification and lowering the human annotation effort. **Novel Class Detection:** Some advanced approaches are developed to explore class emergence problem in data stream like SEEN (Zhu and Li 2020), Zero-Shot-OOD (Tan et al. 2019) and SENC-MaS (Mu et al. 2017) but there is less focus on novel intent detection. **Active Learning:** ModAL (Danka and Horvath 2018) propose various active learning (AL) strategies to improve system accuracy. (Tian and Lease 2011) develop a maximal marginal uncertainty based AL model. But, these are not focused on rejected utterances.

In this work, an end-to-end framework is proposed to identify novel intents with increased in-domain class detection accuracy while handling system rejected utterances.

Dataset

NIDAL approach is compared with similar models and evaluated across several standard public dataset in NLU domain - SNIPS (Coucke et al. 2018), ATIS (Tur, Hakkani-Tür, and Heck 2010) and Facebook Multi-lingual (Schuster et al. 2018). This framework is language agnostic and performs significantly well on all the datasets.

Approach

The framework "NIDAL" is divided into two major components - novel intent detection and active learning. The Novel Intent Detection module detects out-of-domain samples on unlabelled data and the active learning based known intent classification, handles the rejected utterances ensuring higher overall accuracy on the known intents.

Neural Model ($\mathcal{M}$): The JointBERT (Chen, Zhuo, and Wang 2019) is used as the learning model for intent classification.English uncased BERT-Base model is used as the base model for JointBERT. For other languages, bert-base-multilingual-uncased model is used. The best results can be obtained when the model is trained for 10 epochs and the learning rate is $5e-5$.

Part 1: Novel Intent Detection (NID): The algorithm for Novel Intent Detection consists of the following Steps: **Step 1 - *Cycle 0***: With the initial labelled data ($\mathcal{L}$), consisting of examples only from the known intent set, a Neural Model ($\mathcal{M}$) is trained. **Step 2 - *Out-Of-Domain Sample Detection (OOD-SD)***: MSP (Hendrycks and Gimpel 2018) algorithm is considered to detect Out-Of-Domain samples $\mathcal{U}'$ on the Unlabelled Data ($\mathcal{U}$). **Step 3 - *Labeling of Novel Intents***: $m\%$ of the OOD samples is considered $\mathcal{U}'$ for manual annotation ($\mathcal{U}'_m$). Experiments are done with various labeling

Method	FB-EN		FB-ES		FB-TH		ATIS		SNIPS	
	Acc	Mac F1	Acc	Mac F1	Acc	Mac F1	Acc	Mac F1	Acc	Mac F1
SEEN*	84.9	86.2	75.4	79.8	71.9	77.5	84.3	71.4	85.8	86.2
Zero-Shot-OOD*	86.7	87.4	62.1	70.1	68.6	72.1	84.8	81.7	87.4	88.5
SENC-MaS*	82.3	83.9	69.4	70.5	64.6	72.8	80.1	69.4	82.0	81.4
NIDAL	**98.0**	**93.1**	**91.5**	**96.1**	**92.5**	**89.5**	**92.1**	**88.9**	**97.9**	**95.5**

Table 1: Overall Accuracy (Acc) and Macro average F1-score on Facebook-English, Spanish, Thai, ATIS and SNIPS Datasets. * - advantage to these baselines, since they do not distinguish between multiple novel intents.

strategies to add the rest of the OOD samples back to Cycle 0 training set ($\mathcal{L}$). **Step 4 - *Cycle 1***: $\mathcal{M}$ is re-trained on the modified training set to perform prediction on the Test Set.

Part 2: Active Learning (AL): Next, an active learning based framework is devised to take care of rejected utterances with a low confidence. The algorithm is as follows: **Step 1 - *Cycle 1***: The retrained Neural Model ($\mathcal{M}$) is used to predict intent scores on the unlabelled dataset after removing OODs ($\mathcal{U}_{rem}$). **Step 2 - *Cycle 1***: A pre-defined threshold ($\mathcal{TH}$) is set, and the samples with score $< \mathcal{TH}$ are then fed XGBoost (XGB) classifier and the prediction score is used to identify their labels (*auto-corrected*). Samples which are rejected again, are manually annotated. These auto-corrected and annotated samples are added back to the labelled data and removed from the unlabelled data. **Step 3 - *Cycle 2***: The Neural Model ($\mathcal{M}$) is re-trained with updated labelled dataset and redo step 1-2 for the same threshold value. **Step 4:** This approach is run for $2 \leq K \leq 5$ cycles.

Experimental Results

Softmax Prediction Probability (MSP) is used to predict out-of-domain samples based on the softmax prediction scores. Different threshold values on top of Neural Model ($\mathcal{M}$) is set to identify rejected utterances and optimum results are obtained when threshold is set to 75% of maximum classification probability score for a particular dataset. These rejected utterances are passed through XGBoost (XGB) classifier for auto-correction. The rest of the samples (i.e. XGB rejected) are annotated manually. These auto corrected and manually annotated samples are removed from the unlabelled set and added back to labelled dataset, L for retraining purposes. Thus the auto-correction criteria, saves significant human labeling cost.

The results for Novel Intent Detection and Active Learning (NIDAL) in terms of accuracy and macro f1 score for various english (Facebook-English [FB-EN], SNIPS, ATIS) and non-english data (Facebook-Thai [FB-TH] and Facebook-Spanish [FB-ES]) are shown in Table 1.

Competing Baselines: Experiments are done with modified versions of various existing methods to compare NIDAL - SEEN (Zhu and Li 2020), Zero-Shot-OOD (Tan et al. 2019) and SENC-MaS (Mu et al. 2017). Experiments are done with different variations of baselines (parameter and hyperparameter tuning) and the best results are reported. Since, all these methods can only detect one novel class at a time, when considering the multiple novel intents in test set as a single novel class for these methods and report the accuracy and macro-F1 considering the various known classes and a novel class for the same. The results for different baselines are shown in Table 1. It is observed that the annotation cost varies from 6-10% samples of the unlabelled data. Considering the huge gains in performance, this is a reasonable annotation cost trade-off. These experiments show that given the same amount of human labeling, this framework works much better than the baselines.

Conclusion

This paper proposes NIDAL, an end-to-end framework for detection of novel intents, as well as to handle system rejected utterances using an active learning framework. Experiments are performed on various benchmark datasets and the results consistently show the efficacy of the proposed framework. Specifically, it achieves large and consistent gains in performance with a small human annotation cost across different datasets. One future step is to extend this work on diverse domains (like finance, technology, healthcare etc.).

References

Chen, Q.; Zhuo, Z.; and Wang, W. 2019. BERT for Joint Intent Classification and Slot Filling. arXiv:1902.10909.

Coucke, A.; Saade, A.; Ball, A.; et al. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. arXiv:1805.10190.

Danka, T.; and Horvath, P. 2018. modAL: A modular active learning framework for Python. *arXiv preprint arXiv:1805.00979*.

Hendrycks, D.; and Gimpel, K. 2018. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv:1610.02136.

Mu, X.; Zhu, F.; Du, J.; Lim, E.; and Zhou, Z.-H. 2017. Streaming classification with emerging new class by class matrix sketching. In *AAAI Conference*.

Schuster, S.; Gupta, S.; Shah, R.; and Lewis, M. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Tan, M.; Yu, Y.; Wang, H.; Wang, D.; Potdar, S.; Chang, S.; and Yu, M. 2019. Out-of-domain detection for low-resource text classification tasks. *arXiv preprint arXiv:1909.05357*.

Tian, A.; and Lease, M. 2011. Active learning to maximize accuracy vs. effort in interactive information retrieval. In *ACM SIGIR conference*, 145–154.

Tur, G.; Hakkani-Tür, D.; and Heck, L. 2010. What is left to be understood in ATIS? In *2010 IEEE Spoken Language Technology Workshop*, 19–24. IEEE.

Zhu, Y.-N.; and Li, Y.-F. 2020. Semi-Supervised Streaming Learning with Emerging New Labels. In *AAAI Conference*.