Constrained Self-Supervised Clustering for Discovering New Intents (Student Abstract)

Ting-En Lin,[1,2] **Hua Xu,**[1,2*] **Hanlei Zhang**[1,2,3]
[1]State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China,
[2]Beijing National Research Center for Information Science and Technology(BNRist), Beijing 100084, China
[3]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
lte17@mails.tsinghua.edu.cn, xuhua@tsinghua.edu.cn, 16281181@bjtu.edu.cn

Abstract

Discovering new user intents is an emerging task in the dialogue system. In this paper, we propose a self-supervised clustering method that can naturally incorporate pairwise constraints as prior knowledge to guide the clustering process and does not require intensive feature engineering. Extensive experiments on three benchmark datasets show that our method can yield significant improvements over strong baselines.

Introduction

In the dialogue system, it is important to identify new user intents that we do not know beforehand. It can help us to discover new business opportunities and determine the future direction of system development. Since most conversational data are unlabelled, an effective clustering method can help us to find a reasonable taxonomy automatically and understand the pattern of potential user needs. However, it is difficult to get the desired clustering results since the taxonomy of intents is strongly guided by prior knowledge (Lin and Xu 2019). For example, suppose we want to partition the data based on the technical problems encountered by users, we may eventually get clustering results partitioned by question types (e.g., what, how, why). Existing methods incorporate prior knowledge by intensive feature engineering (Shi et al. 2018). It is not only time consuming but also can lead to overfitting.

In a real-world scenario, as shown in Figure 1, we may have limited labeled data and a vast amount of unlabeled data, but we do not know all the intent categories in advance. Besides, the training data is noisy because there are new intents in the unlabeled data. The key is to use limited labeled data to improve clustering performance effectively.

To address the issues, we propose a self-supervised clustering method that leverage labeled data to learn cluster-friendly representations adaptively. Experimental results show that our method significantly improves clustering performance and can even generalize to the new intents that we do not know in advance.

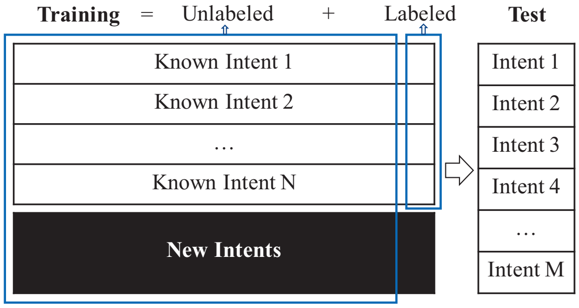

Figure 1: An example of discovering new intents. Our goal is to find out the underlying new intents by utilizing the limited labeled data to aid the clustering process.

Proposed Method

We divide the proposed method into three steps: intent representation, pairwise-classification, and cluster refinement.

Intent Representation We use the pre-trained language model, BERT (Devlin et al. 2019), to obtain intent representations. Given the i^{th} sentence x_i in the corpus, we take all token embeddings $[C, T_1, \cdots, T_N] \in \mathbb{R}^{(N+1)\times H}$ in the last hidden layer of BERT and apply mean-pooling on it to get the representation $e_i \in \mathbb{R}^H$ where N is the sequence length and H is the hidden layer size. Then, we feed e_i to fully connected layer and obtain intent representation $I_i \in \mathbb{R}^k$ where k is the number of clusters.

Pairwise-Classification with Similarity Loss We reframe the clustering problem as a pairwise-classification task. By determining whether the sentence pair is similar or not, our model can learn clustering-friendly intent representation. We use intent representation I to compute the similarity matrix S:

$$S_{ij} = \frac{I_i I_j^T}{\|I_i\| \, \|I_j\|} \tag{1}$$

where $\|\cdot\|$ is L2 norm and $i, j \in \{1, \ldots, n\}$. We denote batch size as n. S_{ij} indicates the similarity the between sentence x_i and x_j. Then, we iteratively go through supervised and self-supervised step to optimize the model.

[*] Hua Xu is the corresponding author.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

For supervised step, we can construct the label matrix R:

$$R_{ij} := \begin{cases} 1, \text{ if } & y_i = y_j, \\ 0, \text{ if } & y_i \neq y_j \end{cases} \tag{2}$$

where $i, j \in \{1, \ldots, n\}$. Then, we use the similarity matrix S and the label matrix R to compute the similarity loss $\mathcal{L}_{\text{sim}}$:

$$\begin{aligned} \mathcal{L}_{\text{sim}}(R_{ij}, S_{ij}) = & -R_{ij} \log(S_{ij}) \\ & -(1 - R_{ij}) \log(1 - S_{ij}). \end{aligned} \tag{3}$$

Here we treat labeled data as prior knowledge and use it to guide the clustering process. It implies how the model should partition the data.

For self-supervised step, we apply dynamic thresholds on similarity matrix S and get the self-labeled matrix $\hat{R}$:

$$\hat{R}_{ij} := \begin{cases} 1, \text{if} \quad S_{ij} > u(\lambda) \quad \text{or} \quad y_i = y_j, \\ 0, \text{if} \quad S_{ij} < l(\lambda) \quad \text{or} \quad y_i \neq y_j, \\ \text{Not selected , otherwise} \end{cases} \tag{4}$$

where $i, j \in \{1, \ldots, n\}$. The dynamic upper threshold $u(\lambda)$ and the dynamic lower threshold $l(\lambda)$ are used to determine whether the sentence pair is similar or dissimilar. Note that the sentence pairs with similarities between $u(\lambda)$ and $l(\lambda)$ do not participate in the training process. In this step, we mix labeled and unlabeled data to train the model.

Then, we use the similarity matrix S and the self-labeled matrix $\hat{R}$ to compute the similarity loss $\hat{\mathcal{L}}_{\text{sim}}$:

$$\begin{aligned} \hat{\mathcal{L}}_{\text{sim}}(\hat{R}_{ij}, S_{ij}) = & -\hat{R}_{ij} \log(S_{ij}) \\ & -(1 - \hat{R}_{ij}) \log(1 - S_{ij}). \end{aligned} \tag{5}$$

We gradually decrease $u(\lambda)$ and increase $l(\lambda)$ to select more sentence pairs to participate in the training process. Please note that it may also introduce more noise to $\hat{R}$. When $u(\lambda) \leq l(\lambda)$, we stop the iterative process. Finally, we refine the cluster assignments with K-means on intent representations I and get the clustering results.

Experiments

We conduct experiments on three publicly available short text datasets: SNIPS, DBPedia, and StackOverflow. Then, we compare our method with both unsupervised and semi-supervised clustering methods. For unsupervised methods, we compare our method with K-means (KM), agglomerative clustering (AG), SAE-KM and DEC, DCN, and DAC (Chang et al. 2017). For semi-unsupervised, we compare with PCK-means, Semi (Wang, Mi, and Ittycheriah 2016) and KCL (Hsu, Lv, and Kira 2018).

For each run of experiments, we randomly select 25% of classes as unknown and 10% of training data as labeled. We set the number of clusters as the ground-truth. We use the same dynamic thresholds as DAC and set $u(\lambda) = 0.95 - \lambda$, $l(\lambda) = 0.455 + 0.1 \cdot \lambda$, and $\eta = 0.009$. We use normalized mutual information (NMI) and clustering accuracy (ACC) as evaluation metrics and report the average performance of each algorithm over ten runs.

The results are shown in Table 1. Our method outperforms baselines by a significant margin on all datasets. It shows that the intent representations learned by pairwise classification and constraints can be effectively grouped into clusters, and can even generalize to the new intents.

Table 1: The clustering results on three datasets. We evaluate both unsupervised and semi-supervised methods.

	SNIPS		DBPedia		StackOverflow	
Method	NMI	ACC	NMI	ACC	NMI	ACC
KM	71.4	84.4	67.3	61.0	8.2	13.6
AG	71.0	75.5	65.6	56.1	10.6	14.7
SAE-KM	78.2	87.9	59.7	50.3	32.6	34.4
DEC	84.6	91.6	53.4	39.6	10.9	13.1
DCN	58.6	57.5	54.5	47.5	31.1	34.3
DAC	80.0	76.3	75.4	64.0	14.7	16.3
PCK-means	74.9	86.9	79.8	83.1	17.3	24.2
KCL	75.2	63.9	83.2	60.6	8.8	13.9
Semi	76.0	78.0	86.4	75.3	65.1	65.3
Ours	**88.0**	**93.0**	**93.4**	**89.8**	**67.7**	**71.5**

Conclusion and Future Work

In this paper, we propose a self-supervised clustering method that leverages limited labeled data to improve the performance of discovering new intents. In the future, we will refine the cluster assignments in an end-to-end fashion.

Acknowledgments This paper is funded by the National Natural Science Foundation of China (Grant No: 61673235) and National Key R&D Program Projects of China (Grant No: 2018YFC1707605).

References

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 5879–5887.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the NAACL-HLT*, 4171–4186.

Hsu, Y.-C.; Lv, Z.; and Kira, Z. 2018. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*.

Lin, T. E., and Xu, H. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5491–5496. Association for Computational Linguistics.

Shi, C.; Chen, Q.; Sha, L.; Li, S.; Sun, X.; Wang, H.; and Zhang, L. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 684–689.

Wang, Z.; Mi, H.; and Ittycheriah, A. 2016. Semi-supervised clustering for short text via deep representation learning. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 31–39.