PSEUDO LABELING AND NEGATIVE FEEDBACK LEARNING FOR LARGE-SCALE MULTI-LABEL DOMAIN CLASSIFICATION

Joo-Kyung Kim *Young-Bum Kim*

Amazon Alexa AI

ABSTRACT

In large-scale domain classification, an utterance can be handled by multiple domains with overlapped capabilities. However, only a limited number of ground-truth domains are provided for each training utterance in practice while knowing as many as correct target labels is helpful for improving the model performance. In this paper, given one ground-truth domain for each training utterance, we regard domains consistently predicted with the highest confidences as additional pseudo labels for the training. In order to reduce prediction errors due to incorrect pseudo labels, we leverage utterances with negative system responses to decrease the confidences of the incorrectly predicted domains. Evaluating on user utterances from an intelligent conversational system, we show that the proposed approach significantly improves the performance of domain classification with hypothesis reranking.

Index Terms— Domain classification, multi-label classification, pseudo labeling, negative feedback learning

1. INTRODUCTION

Domain classification is a task that predicts the most relevant domain given an input utterance [1].[1] It is becoming more challenging since recent conversational interaction systems such as Amazon Alexa, Google Assistant, and Microsoft Cortana support more than thousands of domains developed by external developers [4, 3, 5]. As they are independently and rapidly developed without a centralized ontology, multiple domains have overlapped capabilities that can process the same utterances. For example, "*make an elephant sound*" can be processed by `AnimalSounds`, `AnimalNoises`, and `ZooKeeper` domains.

Since there are a large number of domains, which are even frequently added or removed, it is infeasible to obtain all the ground-truth domains of the training utterances, and domain classifiers for conversational interaction systems are usually trained given only a small number (usually one) of ground-truths in the training utterances. This setting corresponds to multi-label positive and unlabeled (PU) learning, where assigned labels are positive, unassigned labels are not necessarily negative, and one or more labels are assigned for an instance [6, 7].[2]

In this paper, we utilize user log data, which contain triples of an utterance, the predicted domain, and the response, for the model training. Therefore, we are given only one ground-truth for each training utterance. In order to improve the classification performance in this setting, if certain domains are repeatedly predicted with the highest confidences even though they are not the ground-truths of an utterance, we regard the domains as additional pseudo labels. This is closely related to pseudo labeling [8] or self-training [9, 10, 11]. While the conventional pseudo labeling is used to derive target labels for unlabeled data, our approach adds pseudo labels to singly labeled data so that the data can have multiple target labels. Also, the approach is related to self-distillation, which leverages the confidence scores of the non-target outputs to improve the model performance [12, 13]. While distillation methods utilize the confidence scores as the soft targets, pseudo labeling regards high confident outputs as the hard targets to further boost their confidences. We use both pseudo labeling and self-distillation in our work.

Pseudo labels can be wrongly derived when irrelevant domains are top predicted, which can lead the model training with wrong supervision. To mitigate this issue, we leverage utterances with negative system responses to lower the prediction confidences of the failing domains. For example, if a system response of a domain for an input utterance is "`I don't know that one`", the domain is regarded as a negative ground-truth since it fails to handle the utterance.

Evaluating on an annotated dataset from the user logs of a large-scale conversation interaction system, we show that the proposed approach significantly improves the domain classification especially when hypothesis reranking is used [14, 5].

2. MODEL OVERVIEW

We take a hypothesis reranking approach, which is widely used in large-scale domain classification for higher scalabil-

[1] A domain is usually defined as an application or functionality than can handle specific intents [1, 2, 3].

[2] [6] utilizes pairwise label dependencies whose computational complexity is a polynomial of degree 2 in terms of the number of labels, which is unsuitable in large-scale domain classification. [7] is dealing with the scalability issue but they assume optimizing a low-rank linear model whose expressive power is limited.

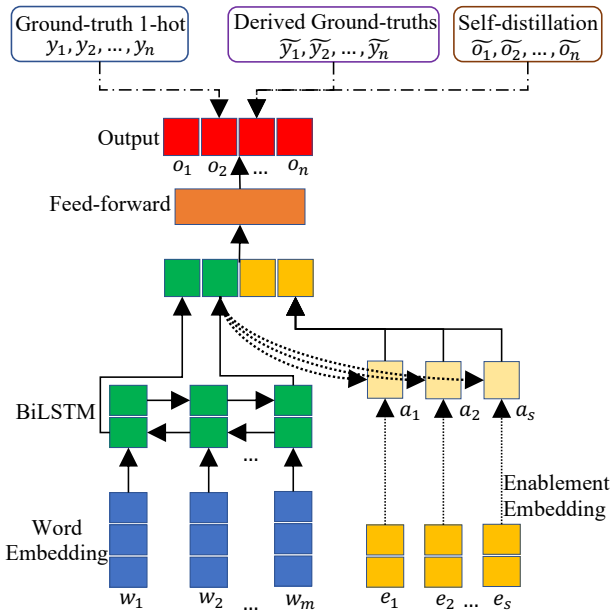

Fig. 1. Shortlister architecture: an input utterance is represented as a concatenation of the utterance vector from BiLSTM and the weighted sum of domain enablement vectors through domain enablement attention mechanism. Then, a feed-forward neural network followed by sigmoid activation represents the n-dimensional output vector.

ity [14, 5]. Within the approach, a shortlister, which is a light-weighted domain classifier, suggests the most promising k domains as the hypotheses. We train the shortlister along with the added pseudo labels, leveraging negative system responses, and self-distillation, which are described in Section 3. Then a hypothesis reranker selects the final prediction from the k hypotheses enriched with additional input features, which is described in Section 4.

3. SHORTLISTER MODEL

Our shortlister architecture is shown in Figure 1. The words of an input utterance are represented as contextualized word vectors by bidirectional long short-term memory (BiLSTM) on top of the word embedding layer [15]. Then, the concatenation of the last outputs of the forward LSTM and the backward LSTM is used to represent the utterance as a vector.[3] Following [3] and [18], we leverage the domain enablement information[4] through attention mechanism [19], where the weighted sum of enabled domain vectors followed by sigmoid activation is concatenated to the utterance vector for representing a personalized utterance. On top of the personalized utterance vector, a feed-forward neural network followed by sigmoid activation is used to obtain n-dimensional output vector o, where the prediction confidence of each domain is represented as a scalar value between 0 and 1.

Given an input utterance and its target label, binary cross entropy is used as the baseline loss function as follows:

$$\mathcal{L}_b = -\sum_{i=1}^{n} y_i \log o_i + (1 - y_i) \log (1 - o_i), \qquad (1)$$

where o, y, and n denote the model output vector, the one-hot vector of the target label, and the number of total labels. We describe other proposed loss functions in the following subsections.

3.1. Deriving Pseudo Labels

We hypothesize that the outputs repeatedly predicted with the highest confidences are indeed correct labels in many cases in multi-label PU learning setting. This approach is closely related to pseudo labeling [8] or self-training [9, 10, 11] in semi-supervised learning since our model is supervised with additional pseudo labels, but differs in that our approach assigns pseudo labels to singly labeled train sets rather than unlabeled data sets.

We derive the pseudo labels when the following conditions are met:

- Maximally p domains predicted with the highest confidences that are higher than the confidence of the known ground-truth.
- Domains predicted with the highest confidences for r times consecutively so that consistent top predictions are used as pseudo labels.

For the experiments in Section 5, we use p=2 and r=4, which show the best dev set performance. Those derived pseudo labels are used in the model training as follows:

$$\mathcal{L}_d = -\sum_{i=1}^{n} \tilde{y}_i \log o_i + (1 - \tilde{y}_i) \log (1 - o_i), \qquad (2)$$

where $\tilde{y}$ denotes an n-hot vector such that the elements corresponding to the original ground-truth and the additional pseudo labels are set to 1.

3.2. Leveraging Negative Feedback

During the model training, irrelevant domains could be top predicted, and regarding them as additional target labels results in wrong confirmation bias [20], which causes incorrect model training. To reduce the side effect, we leverage utterances with negative responses in order to discourage the utterances' incorrect predictions. This setting can be considered as a multi-label variant of Positive, Unlabeled, and Biased Negative Data (PUbN) learning [21].

We obtain training utterances from log data, where utterances with positive system responses are used as the positive train set in Equation 1 and 2 while the utterances with negative responses are used as the negative train set in Equation 3.

[3] In our experiments, using convolution neural networks [16] or self-attention [17] for encoding do not make significant differences.

[4] Enabled domains are favorite or authenticated domains.

For example, `AnimalSounds` is a (positive) ground-truth domain for "*a monkey sound*" because the system response to the utterance is "*`Here comes a monkey sound`*" while it is a negative ground-truth for "*a dragon sound*" as the response is "*`I don't know what sound a dragon makes`*".[5] [6]

Previous work [22, 23] excludes such negative utterances from the training set. We find that it is more effective to explicitly demote the prediction confidences of the domains resulted in negative responses if they are top ranked. It is formulated as a loss function:

$$\mathcal{L}_n = \begin{cases} -\log(1-o_j) & \forall i \neq j \; o_i \leq o_j \\ 0 & \text{Otherwise,} \end{cases} \quad (3)$$

where j denotes the index corresponding to the negative ground-truth domain. We demote the confidences of the negative ground-truths only when they are the highest so that the influence of using the negative ground-truths is not overwhelming.[7]

3.3. Self-distillation

Knowledge distillation has been shown to improve the model performance by leveraging the prediction confidence scores from another model or from previous epochs [12, 13, 18]. Inspired by [18], we utilize the model at the epoch showing the best dev set performance before the current epoch to obtain the prediction confidence scores as the soft target. The self-distillation in our work can be formulated as follows:

$$\mathcal{L}_s = -\sum_{i=1}^{n} \tilde{o_i} \log o_i + (1-\tilde{o_i}) \log(1-o_i), \quad (4)$$

where $\tilde{o_i}$ denotes the model output at the epoch showing the best dev set performance so far. Before taking sigmoid to obtain $\tilde{o_i}$, we use 16 as the temperature to increase the influence of distillation [12], which shows the best dev set performance following [18].

3.4. Combined Loss

The model is optimized with a combined loss function as follows:

$$\mathcal{L} = \left(1-\alpha^t\right)\mathcal{L}_b + \alpha^t\left(\mathcal{L}_d + \mathcal{L}_s\right) + \beta\mathcal{L}_n. \quad (5)$$

where $\alpha^t = 1 - 0.95^t$ and t is the current epoch so that the baseline loss is mainly used in the earlier epochs while the pseudo labels and self-distillation are more contributing in the later epochs following [24]. β is a hyperparameter for utilizing negative ground-truths, which is set to 0.00025 showing the best dev set performance.

[5] Recent intelligent conversational systems can support thousands of domains, many of which are with very specific/narrow capabilities. For "*a dragon sound*", `DungeonSound` and `DragonFire` are the correct domains, and predicting `ZooKeeper` is incorrect.

[6] Negative responses can be easily identified since the responses are generated from predefined templates. We use 2K template patterns to extract such responses.

[7] In our experiments, reducing confidences of negative ground-truths regardless of the confidence ranks shows worse performance.

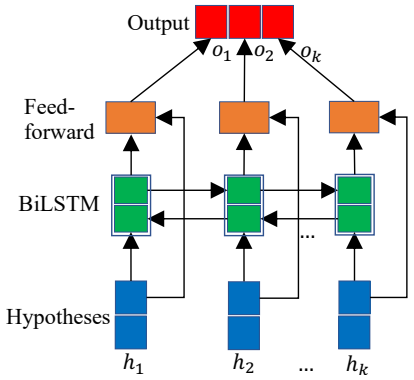

Fig. 2. Hypothesis Reranker Architecture: Each hypothesis consists of scores and vectors of domain, intent, and slots. Then, BiLSTM and a feed-forward neural network are used to represent contextualized hypothesis confidence scores.

4. HYPOTHESIS RERANKING MODEL

Figure 2 shows the overall architecture of the hypothesis reranker that is similar to [5]. First, we run intent classification and slot filling for the k most confident domains from the shortlister outputs to obtain additional information for those domains [1].[8] Then, we compose k hypotheses, each of which is a vector consists of the shortlister confidence score, intent score, Viterbi score of slot-filling, domain vector, intent vector, and the summation of the slot vectors. On top of the k hypothesis vectors, a BiLSTM is utilized for representing contextualized hypotheses and a shared feed-forward neural network is used to obtain final confidence score for each hypothesis. We set k=3 in our experiments following [5]. We leverage the given ground-truth and the derived pseudo labels from the shortlister at the epoch showing the best dev set performance as target labels for training the reranker. We use hinge loss with margin 0.4 as the loss function.

One issue of the hypothesis reranking is that a training utterance cannot be used if no ground-truth exist in the top k predictions of the shortlister. This is problematic in the multi-label PU setting since correct domains can indeed exist in the top k list but unknown, which makes the training utterance less useful in the reranking. Our pseudo labeling method can address this issue. If correct pseudo labels are derived from the shortlister's top predictions for such utterances, we can use them properly in the reranker training, which was unavailable without them. This allows our approach make more improvement in hypothesis reranking than shortlisting.

[8] Maximum Entropy model and Conditional Random Field model [25] are utilized for intent classification and slot-filling, respectively.

	Model	Shortlister				Hypothesis Reranker		
		Precision	Recall	F-1	$nDCG_3$	Precision	Recall	F-1
(1)	Base	77.27	**83.27**	80.15	71.92	79.13	82.34	80.71
(2)	Base+pseudo	76.77	82.87	79.70	71.64	79.02	81.23	80.11
(3)	Base+neg_feed	77.90	81.32	79.58	72.15	79.33	83.54	81.38
(4)	Base+neg_feed+self_dist	77.73	82.46	80.03	72.24	79.24	83.71	81.41
(5)	Base+pseudo+neg_feed	**78.14**	82.87	80.43	72.53	**79.52**	83.89	81.65
(6)	Base+pseudo+neg_feed+self_dist	77.96	83.21	**80.50**	**72.68**	79.41	**84.09**	**81.69**

Table 1. Evaluation results on various metrics (%). pseudo, neg_feed, and self_dist denote using derived pseudo labels, negative feedback, and self-distillation, respectively.

Utterance	Known ground-truth	Additional pseudo labels
One hundred twenty beats per minute	Acoustic Metronome	My Metronome, Metronome Lite
Play ocean sounds	Ambient Sounds	Sleep and Relaxation Sounds, Sleep Sounds: Ocean Sounds
Give me the news briefing	CBS News	The Washington Post, CNN

Table 2. Examples of additional pseudo labels.

5. EXPERIMENTS

In this section, we show training and evaluation sets, and experiment results.

5.1. Datasets

We utilize utterances with explicit invocation patterns from an intelligent conversational system for the model training similarly to [5] and [18]. For example, given "ask {`AmbientSounds`} to {*play thunderstorm sound*}", we extract "*play thunderstorm*" as the input utterance and `Ambient Sounds` as the ground-truth. One difference from the previous work is that we utilize utterances with positive system responses as the positive train set and the dev set, and use those with the negative responses as the negative train set as described in Section 3.2. We have extracted 3M positive train, 400K negative train, and 600K dev sets from 4M log data with 2,500 most frequent domains as the ground-truths. Pseudo labels are added to 53K out of 3M in the positive train set as described in Section 3.1.

For the evaluation, we have extracted 10K random utterances from the user log data and independent annotators labeled the top three predictions of all the evaluated models for each utterance so that we can correctly compute nDCG at rank position 3.

5.2. Experiment Results

Table 1 shows the evaluation results of the shortlister and the hypothesis reranker with the proposed approaches. For the shortlisters, we show $nDCG_3$ scores, which are highly correlated with the F1 scores of the rerankers than other metrics since the second and third top shortlister predictions contribute the metric. We find that just using the pseudo labels as the additional targets degrades the performance (2). However, when both the pseudo labels and the negative ground-truths are utilized, we observe significant improvements for both precision and recall (5). In addition, recall is increased when self-distillation is used, which achieves the best F1 score (6). Each of utilizing the negative feedback ($(1) \rightarrow (3)$ and $(2) \rightarrow (5)$) and then additional pseudo labels ($(3) \rightarrow (5)$ and $(4) \rightarrow (6)$) show statistically significant improvements with McNemar test for p=0.05 for the final reranker results.

Using self-distillation ($(3) \rightarrow (4)$ and $(5) \rightarrow (6)$) shows increased F-1 score by increasing recall and decreasing precision, but the improvements are not significant. One issue is that pseudo labeling and self-distillation are contrary since the former encourages entropy minimization [26, 8] while the latter can increase entropy by soft targeting the non-target labels. More investigation of self-distillation along with the proposed pseudo labeling would be future work.

Table 2 shows examples of derived pseudo labels from model (6). It demonstrates that the domains capable of processing the utterances can be derived, which helps more correct model training.

6. CONCLUSION

We have proposed deriving pseudo labels along with leveraging utterances with negative system responses and self-distillation to improve the performance of domain classification when multiple domains are ground-truths even if only one ground-truth is known in large-scale domain classification. Evaluating on the test utterances with multiple ground-truths from an intelligent conversational system, we have showed that the proposed approach significantly improves the performance of domain classification with hypothesis reranking.

As future work, combining our approach with pure semi-supervised learning, and the relation between pseudo labeling and distillation should be further studied.

7. REFERENCES

[1] Gokhan Tur and Renato de Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, New York, NY: John Wiley and Sons, 2011.

[2] Ruhi Sarikaya, Paul A Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Celikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, and Xiaohu Liu, "An overview of end-to-end language understanding and dialog management for personal digital assistants," in *SLT*, 2016, p. 391–397.

[3] Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, and Ruhi Sarikaya, "Efficient Large-Scale Neural Domain Classification with Personalized Attention," in *ACL*, 2018, pp. 2214–2224.

[4] Anjishnu Kumar, Arpit Gupta, Julian Chan, Sam Tucker, Bjorn Hoffmeister, Markus Dreyer, Stanislav Peshterliev, Ankur Gandhe, Denis Filiminov, Ariya Rastrow, Christian Monson, and Agnika Kumar, "Just ASK: Building an Architecture for Extensible Self-Service Spoken Language Understanding," in *NIPS Workshop on Conversational AI*, 2017.

[5] Young-Bum Kim, Dongchan Kim, Joo-Kyung Kim, and Ruhi Sarikaya, "A scalable neural shortlisting-reranking approach for large-scale domain classification in natural language understanding," in *NAACL*, 2018, pp. 16–24.

[6] Atsushi Kanehira and Tatsuya Harada, "Multi-label ranking from positive and unlabeled data," in *CVPR*, 2016, pp. 5138–5146.

[7] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon, "Large-scale multi-label learning with missing labels," in *ICML*, 2014, pp. 593–601.

[8] Dong-Hyun Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," in *ICML WREPL*, 2013.

[9] David Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *ACL*, 1995, pp. 189–916.

[10] David McClosky, Eugene Charniak, and Mark Johnson, "Reranking and Self-training for Parser Adaptation," in *ACL*, 2006, pp. 337–344.

[11] Sebastian Ruder and Barbara Plank, "Strong Baselines for Neural Semi-Supervised Learning under Domain Shift," in *ACL*, 2018, pp. 1044–1054.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning Workshop*, 2014.

[13] Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar, "Born Again Neural Networks," in *ICML*, 2018, pp. 1602–1611.

[14] Jean-Philippe Robichaud, Paul A. Crook, Puyang Xu, Omar Zia Khan, and Ruhi Sarikaya, "Hypotheses ranking for robust domain classification and tracking in dialogue systems," in *Interspeech*, 2014, pp. 145–149.

[15] Alex Graves and Jürgen Schmidhuber, "Frame-wise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[16] Yoon Kim, "Convolutional neural networks for sentence classification," in *EMNLP*, 2014, pp. 1292–1302.

[17] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, and Yoshua Zhou, Bowen and-Bengio, "A structured self-attentive sentence embedding," in *ICLR*, 2017.

[18] Joo-Kyung Kim and Young-Bum Kim, "Supervised Domain Enablement Attention for Personalized Domain Classification," in *EMNLP*, 2018, pp. 894–899.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[20] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017, pp. 1195–1204.

[21] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama, "Classification from Positive, Unlabeled and Biased Negative Data," in *ICML*, 2019, pp. 2820–2829.

[22] Jason Weston, "Dialog-based language learning," in *NIPS*, 2016, pp. 829–837.

[23] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston, "Learning from dialogue after deployment: Feed yourself, chatbot!," in *ACL*, 2019, pp. 3667–3676.

[24] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing, "Harnessing Deep Neural Networks with Logic Rules," in *ACL*, 2016, pp. 2410–2420.

[25] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[26] Yves Grandvalet and Yoshua Bengio, "Semi-supervised learning by entropy minimization," in *NIPS*, 2005, pp. 529–536.