X-SHOT: LEARNING TO RANK VOICE APPLICATIONS VIA CROSS-LOCALE SHARD-BASED CO-TRAINING

Zheng Gao, Mohamed AbdelHady, Radhika Arava, Xibin Gao, Qian Hu, Wei Xiao, Thahir Mohamed

Amazon Alexa AI, Seattle, USA

ABSTRACT

Virtual assistants such as Google Assistant and Amazon Alexa host thousands of voice applications (skills) that handle a very large and diverse array of customer utterances. However, the number of supported skills may be much lower in some locales, particularly in countries other than the United States. Accordingly, customer utterances handled in a popular locale may be going unclaimed in another locale. Moreover, locales with smaller skill ecosystems also suffer from limited labeled data for training systems to route utterances to skills. To tackle these aforementioned challenges, we propose a Cross-locale SHard-based cO-Training model (X-SHOT) that uses an iterative label augmentation approach to retrieve relevant skills in a source locale for unclaimed utterances in a target locale. The obtained results could be further used by skill developers in the source locale to gauge the latent demand for their skills in other locales and therefore to prioritize the internationalization of their skills accordingly.

Index Terms— Co-Training, Transfer Learning, Pseudo Labeling, Semi-Supervised Learning

1. INTRODUCTION

In Spoken Language Understanding (SLU) system of virtual assistants such as Google Assistant and Amazon Alexa, a skill refers to a third-party voice application created by external developers and used to respond to customer utterances. Skill ranking is the associated task to retrieve the most relevant skills for customer utterances [1]. For example, utterance "*alexa, play today's hits*" will directly invoke skill "*Pandora*" to play trending music in the device.

However, modern SLU systems are faced with two severe challenges. **First**, there is a lack of developed skills in certain locales. A locale is defined as a country with a specific language, for example locale en-US contains all English utterances of the United States. Although SLU systems simultaneously support mono-locale skill ranking in each individual locale, skills are rarely shared across locales. Especially, newly served locales may only support few skills. **Second**, there is a lack of labeled data. For each utterance, we can only invoke the most relevant skill and receive its explicit label ("*positive*" or "*negative*") from customer feedback, while most of skill labels remain unknown.

Cross-domain investigations can naturally solve target locale retrieval tasks by transferring source locale knowledge via adversarial training [2] or knowledge distillation [3, 4] techniques. However they can neither enlarge the scope of skill candidates (i.e. retrieving new skills unavailable in target locale) nor deal with unlabeled data. Other investigations generate pseudo labels on unlabeled data via Positive-Unlabeled (PU) learning [5] or self-training [6] techniques. However they only explore within-locale data and bring in no new knowledge from external resources.

Unclaimed utterances with no suggested skills might be caused by that their appropriate skills are not yet supported in the target locale. One solution is to build a fallback skill retrieval system that can find potential skills in another locale to handle such unclaimed utterances. Based on these, we propose a Cross-locale SHard-based cO-Training (X-SHOT) model by treating locale-specific knowledge as different views to retrieve source locale skills for target locale unclaimed utterances. To alleviate cross-lingual problem, two locales are with the same language. Although the retrieved source locale skills can't immediately take effect on unclaimed utterances because of their absence in target locale, we can track and accumulate their traffic across time and suggest skill developers to enable the top ranked skills with high-volume in target locale.

Our proposed X-SHOT model is a two-step approach with Shortlisting and Reranking. The Shortlisting step performs keyword-based matching to select the best relevant skills. The Reranking step first splits whole cross-locale utterances equally and horizontally into shards, then incrementally trains two locale-specific Reranker models on utterance Shortlisting results in each data shard. The contribution of this work is threefold: **First**, unlike previous works to just regard cross-locale knowledge as auxiliary information for target locale skill retrieval, our proposed X-SHOT model is a step further to retrieve source locale skills as well. **Second**, a combination of biased upsampling and co-training strategy is particularly designed for pseudo labeling, where the former emphasizes the importance of existing positive labels (label exploitation) and the latter extends to more potentially relevant skills (label exploration). **Third**, evaluation results on two real-world datasets certify our model superiority compared with several strong alternatives on automated and human metrics.

2. METHOD

2.1. Model Formulation

Algorithm 1 X-SHOT Model

Input: cross-locale utterances $U = \{U^s, U^t\}$, ground truth skill labels $Y = \{Y^s, Y^t\}$, source locale Elasticsearch index E^s, target locale Elasticsearch index E^t, Elasticsearch skill length K, number of shards N, biased upsampling factor α;
Initialization: source locale Reranker R^s, target locale Reranker R^t;
Split $\{U, Y\}$ into N equal data shards;
$i = 0$;
while $i < N$ **do**
  Current data shard utterance $U_i = \{U_i^s, U_i^t\}$;
  For U_i^s, retrieve its top K Shortlisting skills from E^s as V_s^s, and top K Shortlisting skills from E^t as V_t^s;
  For U_i^t, retrieve its top K Shortlisting skills from E^s as V_s^t, and top K Shortlisting skills from E^t as V_t^t;
  Apply biased upsampling on labeled utterances $\in U_i$ with upsampling factor α;
  if $i > 0$ **then**
    $L_s^s \leftarrow$ positive skill labels predicted by R^s on V_s^s;
    $L_t^s \leftarrow$ positive skill labels predicted by R^t on V_t^s;
    $L_s^t \leftarrow$ positive skill labels predicted by R^s on V_s^t;
    $L_t^t \leftarrow$ positive skill labels predicted by R^t on V_t^t;
    $Y_i^s \leftarrow Y_i^s + (L_s^s \cap L_t^s) \in V_s^s$;
    $Y_i^t \leftarrow Y_i^t + (L_s^t \cap L_t^t) \in V_t^t$;
  end
  Incrementally train R^s with V_s^s, Y_i^s, U_i^s;
  Incrementally train R^t with V_t^t, Y_i^t, U_i^t;
  $i \leftarrow i + 1$
end

In this paper, we aim to retrieve source locale skills to target locale unclaimed utterances. If unique source locale skills are retrieved, our model extends the scope of skills; if common skills having already developed in target locale are retrieved, our model reinforces the skill retrieval capability by involving external locale knowledge. Conventional pointwise model $f_\theta(\cdot)$ aim to minimize the cross-entropy discrepancy between predicted label $f_\theta(u, v)$ and ground truth binary label $y \in \{0, 1\}$ for each pair of utterance u and skill v:

$$\underset{\theta}{\operatorname{argmin}} -y \log f_\theta(u, v) - (1 - y) \log(1 - f_\theta(u, v)) \quad (1)$$

However pointwise models are always hard to converge. Instead, we propose a two-step listwise approach to firstly retrieve the top K most relevant skills, then co-train two locale-specific skill Reranker models (in Algorithm 1). In this way, we only need to minimize the prediction discrepancy for each utterance u and its filtered Shortlisting skill sequence V where Y is its ground truth label sequence:

$$\underset{\theta}{\operatorname{argmin}} \sum_{v \in V, y \in Y} -y \log f_\theta(u, v) - (1 - y) \log(1 - f_\theta(u, v)) \quad (2)$$

Evolving from this, $U = \{U^s, U^t\}$ is extended to represent the training utterances for source locale s and target locale t with ground truth skill labels $Y = \{Y^s, Y^t\}$. The whole data $\{U, Y\}$ are split equally into N shards. For each utterance in the i_{th} data shard, we retrieve its top K most relevant skills from both locales (*Section 2.2*). A data augmentation approach is applied on Shortlisting skills with a combination of biased upsampling and pseudo labeling (*Section 2.3*). The augmented data is used to incrementally train Reranker models R^s and R^t (*Section 2.4*) where unknown/rejected skills are labeled as "negative". Figure 1 is a snippet to visualize how X-SHOT model is trained in the i_{th} target locale data shard, same as source locale.

The X-SHOT model has two advanced characteristics: **First**, the pseudo labeling approach integrates the retrieved skills from both locales, which brings more signals and is usually more reliable than single locale retrieved skills. **Second**, in each iteration, unlike other self-training approaches to keep predicting on the same unlabeled data, our model always trains and predicts on a new data shard, which keeps absorbing new knowledge from both locales.

2.2. Shortlisting

Let $u^a \in U_i$ represents an utterance in the i_{th} data shard belonging to either source or target locale with $a \in \{s, t\}$. The Shortlisting step aims to retrieve its top K most relevant skills in each locale, which can filter out most of irrelevant skills and significantly reducing the workload of subsequent reranking process. We empirically apply TF-IDF model [7] within Elasticsearch for Shortlisting skill candidate generation. Source locale Elasticsearch index E^s is constructed by the descriptions of all source locale skills D^s; target locale Elasticsearch index E^t is constructed by the descriptions of all target locale skills D^t. The similarity score for each pair of utterance u^a and skill v in locale $b \in \{s, t\}$ is calculated as:

$$score(u^a, v|b) = \sum_{w \in u^a} \frac{N_{wv}}{|v|} \log \frac{|D^b|}{|\{j : w \in D_j^b\}|} \quad (3)$$

For each word w in the utterance u^a, N_{wv} denotes the count of word w appeared in description of skill v; $|v|$ denotes the description length of skill v; $|D^b|$ denotes the number of skills in Elasticsearch index E^b; $|\{j : w \in D_j^b\}|$ denotes the number of skill descriptions containing word w in locale b. In the end, we calculate all utterance-skill pairwise scores in both locales via E^s and E^t, and retrieve two associated top K skill sequences V_s^a and V_t^a ranked by similarity scores.

2.3. Data Augmentation

To obtain more labeled data for training, two strategies are jointly applied on each cross-locale data shard including biased upsampling and pseudo labeling. Biased upsampling reinforces the exploitation on utterances with positive labeled skills, and pseudo labeling explores other potentially relevant skills for utterances. As pseudo labels predicted by Reranker models are less reliable because of prediction errors, biased upsampling is applied first for positive label augmentation.

2.3.1. Biased Upsampling

To emphasize the impact of labeled data where positive skill labels are in the utterance Shortlister skill sequences, we sample αT^a labeled utterances in locale $a \in \{s, t\}$ and add them back to original data, where $\alpha \in \mathbb{N}^+$ is the upsampling scale factor and T^a is the number of labeled utterances. The biased upsampling intentionally assigns higher sampling probabilities on the utterances with less likely labeled skills. For a labeled utterance u^a, its sampling probability is calculated as:

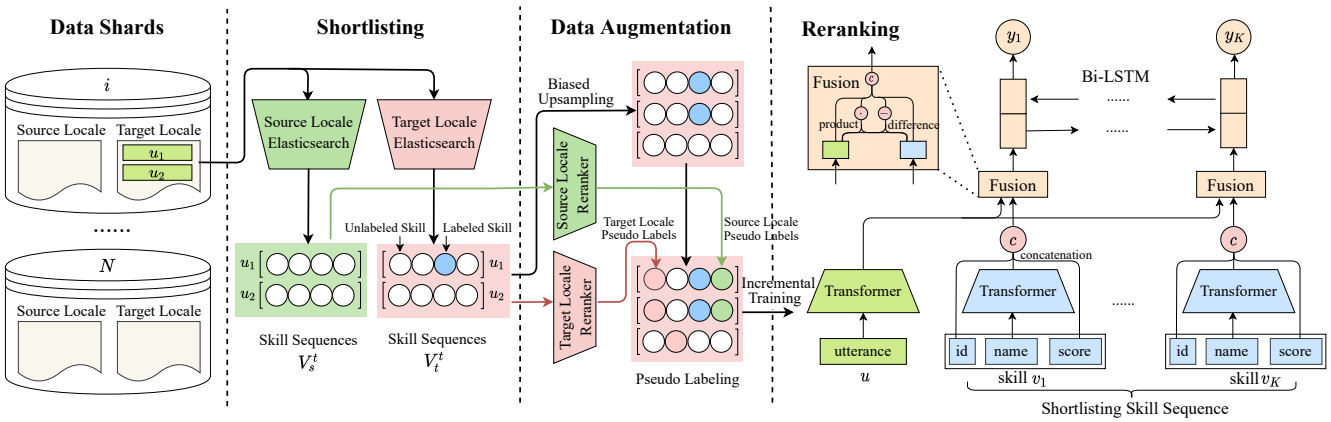


Fig. 1: X-SHOT model training in target locale. This figure shows how target locale Reranker model R^t is updated in the i_{th} data shard. The same training process happens in the source locale to train Reranker model R^s.

$$p(u^a|a \in \{s,t\}) = \frac{|v_u^a|^\varphi}{\sum_j |v_j^a|^\varphi} \tag{4}$$

where v_u^a is the labeled skill of u^a. Exponent $\varphi \in (0,1)$ is 0.75 according to [8]. $|v_u^a|$ is the count of this labeled skill in locale a of current data shard, which is normalized by the sum of all utterances' labeled skill scores.

2.3.2. Pseudo Labeling

In each data shard, an utterance u^a in locale $a \in \{s,t\}$ can obtain two skill sequences V_s^a and V_t^a from Elasticsearch index E^s and E^t. The pseudo labels for u^a are generated from Shortlisting results of both locales. Specifically, Reranker R^a is applied to predict positive labels L_a^a for V_a^a; Reranker R^b is applied to predict positive labels L_b^a for V_b^a where $b = \{s,t\} \setminus a$. In the end, u^a receives two pseudo skill label sets L_a^a and L_b^a carrying on both locale knowledge. We empirically choose their intersective labels $(L_a^a \cap L_b^a) \in V_a^a$ (details in Section 3.7) to add to ground truth labels Y_a^a.

The pseudo labeling approach is applied in both locales simultaneously. To keep a rigorous pseudo label selection throughout the whole training process, two additional labeling criteria are employed: **First**, because of the trustworthy customer feedback, the labels of rejected skills will always remain negative. **Second**, the prediction capacity of Reranker models keep growing with more data shards involved. To ensure high quality pseudo labels, we heuristically designs an adaptive threshold $1 - \frac{p}{1+\ln(i)}$ to filter pseudo labels in the i_{th} data shard where $p \in (0, 0.5)$ denotes the uncertainty rate.

2.4. Reranking

In both locales, Reranker models R^s and R^t take utterances and their Shortlisting skill sequences as model input to predict their skill label sequences.

2.4.1. Utterance and Skill Encoder

We consider three types of information for each skill $v \in V_a^a$, including skill id v_{id}, skill name v_{na} and skill Shortlisting score bin v_{sc}. Transformer [9] is applied with its first hidden state h_u as utterance u^a vector representation with byte-pair-encoding bpe$(\cdot)$ [10] encoding and position embedding pos$(\cdot)$. Skill name is encoded by Transformer in the same way as h_{na}. The concatenation of v_{id}, h_{na} and v_{sc} forms skill vector representation h_s.

$$\begin{aligned} h_u &= \text{Transformer}(\text{bpe}(u^a), \text{pos}(u^a)) \\ h_{na} &= \text{Transformer}(\text{bpe}(v_{na}), \text{pos}(v_{na})) \\ h_s &= [v_{id}; h_{na}; v_{sc}] \end{aligned} \tag{5}$$

2.4.2. Fusion Layer

For each skill $v \in V_a^a$, a fusion layer is proposed to measure its relevance with the target utterance u^a. Besides the original vectors, we also calculate their element-wise product and absolute difference with one layer transformation.

$$\begin{aligned} e_u &= \text{ReLU}(W_u h_u + b_u) \\ e_s &= \text{ReLU}(W_s h_s + b_s) \\ e &= \text{Dropout}([e_u; e_s; e_u \cdot e_s; |e_u - e_s|]) \end{aligned} \tag{6}$$

Where Ws and bs are related weights and bias for each layer. ReLU$(\cdot)$ activiation function and dropout mechanism are both applied. In the end, an interaction vector e is learned for utterance u^a and each skill v as the input of each decoding step.

2.4.3. Sequence Decoder

The interaction vectors of whole skill sequence $E = \{e_1, ..., e_K\}$ is used to predict the ground truth skill label sequence $Y_a^a = \{y_1, ..., y_K\}$. As skills in the Shortlisting skill sequence are not isolated with each other, Bi-LSTM is used to capture skill dependencies. Its i_{th} step output o_i is used to predict the related skill label $\hat{y}_i$.

$$\begin{aligned} o_i &= [\overrightarrow{o}_i; \overleftarrow{o}_i] \\ &= \text{Bi-LSTM}(e_i, \overrightarrow{o}_{i-1}, \overleftarrow{o}_{i+1}) \\ \hat{y}_i &= \sigma(W_o o_i + b_o) \end{aligned} \tag{7}$$

Same as Eq. 2, the final loss is the sequence cross entropy loss $\mathcal{L}$ to minimize the difference between predicted labels $\hat{Y}_a^a$ and ground truth labels Y_a^a:

$$\mathcal{L} = \sum_{y \in Y_a^a, \hat{y} \in \hat{Y}_a^a} -y \log \hat{y} - (1-y) \log(1-\hat{y}) \tag{8}$$

2.5. Training Strategy

Three training tricks are employed to improve model efficiency and efficacy: **First**, the Reranking step is trained incrementally in each data shard instead of trained from scratch, which continuously learns new knowledge from both locales. **Second**, pretrained BERT model is applied for initialization. **Third**, an early stopper is applied to halt model training if performance on validation data drops consecutively.

3. EXPERIMENTS

3.1. Datasets

Two real-world cross-locale datasts from Alexa are constructed for model evaluation. Dataset US-CA takes the United States for source locale (SL) and Canada for target locale (TL). Dataset US-GB takes the United States for source locale and Great Britain for target locale. Both datasets are sampled English utterances (in Table 1).

Dataset	Type	#Utt.	#Skill	%Common Skill	%Labeled
US-CA	SL Training	1,073,678	63,013	36.06%	35.14%
	TL Training	78,146	23,672	95.98%	40.01%
	Testing	11,943	37,490	44.61%	100%
	Validation	8,217	29,476	42.54%	100%
	Inference	222,948	62,079	36.51%	-
US-GB	SL Training	1,073,123	63,012	42.64%	35.27%
	TL Training	415,431	32,952	81.54%	37.70%
	Testing	77,452	52,763	46.43%	100%
	Validation	8,772	29,336	44.60%	100%
	Inference	1,073,132	62,381	42.36%	-

Table 1: Statistics of two cross-locale datasets.

In each dataset, 20% of common skills already developed in both locales are randomly selected as testing skills. The selected testing skills are removed from target locale to simulate the scenario that retrieved skills should come from source locale. The remaining utterances in target local construct TL training data and the whole source locale utterances construct SL training data. Inference data are target locale unclaimed utterances in production. %Labeled calculates the ratio of utterances having positive skills. Positive skills of inference utterances are unknown.

3.2. Baselines

Six baselines are chosen from either skill ranking (Elasticsearch, Pointwise, Listwise) or pseudo labeling (Upsampling, PU learning, Relabeling) perspective: 1) **Elasticsearch** [11]: A search engine use TF-IDF to retrieve the Top-1 relevant skill. 2) **Pointwise** [12]:Neural Collaborative Filtering (NCF) predicts the binary label of each utterance and skill pair. 3) **Listwise** [13]: Bi-LSTM with CRF layer predicts the sequence labels for each utterance on its associated top Elasticsearch skills. 4) **Upsampling**: Upsample positively labeled data to reach equal number of unlabeled data for Listwise model training. 5) **PU learning** [14]: Positive-Unlabeled learning model trains Listwise model on labeled utterances and predicts on unlabeled utterances to generate augmented pseudo labels. 6) **Relabeling**: FAISS model [15] retrieves utterance neighbors to provide pseudo labeled data to train Listwise model. All baselines are trained on the combined locale utterances to have a fair comparison.

3.3. Evaluation Metrics

Three types of evaluation metrics are considered in this paper, including classification metrics, ranking metrics and human metrics. The first two types of automated metrics are for testing data, the human metrics are for inference data. 1) **Classification metrics:** Precision_a calculates the ratio of correct positive predictions. Precision_b calculates the number of correct positive predictions over all positive predictions where the related utterances are with ground truth accepted skills. Their associated F1 scores F1_a, F1_b and Recall are reported as well. 2) **Ranking metrics:** Ranking metrics are reported based on confidence scores of all predicted positive skills, including Precision, Recall, F1, NDCG and Mean Average Precison (MAP). 3) **Human metrics:** In each cross-locale dataset, 400 inference utterances are randomly sampled. Then all their predicted positive skills are integrated from seven model predictions (six baselines + X-SHOT model). Binary labels are manually annotated to each pair of utterance and retrieved skill to calculate Precision, Recall and F1 score.

3.4. Model Comparison

Due to Alexa customer data confidential policy, we are not allowed to directly report their absolute metric scores. Instead, Table 2 shows the normalized performance difference between each baseline and our X-SHOT model, which is calculated as the performance difference between baseline and X-SHOT model, divided by X-SHOT model performance. It reflects how much the baseline models outperform/underperform our proposed model.

In Table 2, for classification metrics, Elasticsearch performs the worst in both datasets, revealing the necessity of Reranking step for fine-grained training. Moreover, three baselines (Listwise, Upsampling, and PU learning) all achieve higher precisions in US-CA dataset than our model. For ranking metrics, the X-SHOT model beats all baselines with around 30%-80% improvement, meaning our model is capable to retrieve adequate number of skills to utterances with appropriate ranking sequence. For human metrics, Elasticsearch surprisingly achieves the best performance in recall and F1 score but sacrifices a lot in precision. One plausible interpretation is that relevant skills tend to rank high in Elasticsearch skill list, certifying the Shortlisting usefulness.

3.5. Model Prediction Statistics

To further explore the prediction results in both datasets, we summarize several prediction statistics of testing and inference data in Table 3. %Retrieval calculates the ratio of utterances with retrieved skills. Testing data in both datasets have higher retrieval rate (%Retrieval) and average retrieved skills (#Label/Utt.) compared with inference data, meaning the X-SHOT model is more competent to rank skills for testing utterances. Moreover, %Common Skills is the rate of re-

Dataset	Model	Classification Metrics					Ranking Metrics					Human Metrics		
		Recall	Precision$_a$	F1$_a$	Precision$_b$	F1$_b$	Precision	Recall	F1	NDCG	MAP	Precision	Recall	F1
US-CA	Elasticsearch	-64.76%	-56.79%	-59.45%	-47.79%	-56.28%	-58.76%	-63.72%	-60.37%	-63.65%	-61.21%	-79.45%	**+260.08%**	**+33.54%**
	Pointwise	-70.81%	-19.55%	-46.43%	-31.87%	-51.37%	-64.68%	-69.83%	-66.34%	-62.18%	-63.91%	-20.00%	-60.08%	-57.87%
	Listwise	-53.99%	**+71.76%**	-4.34%	**+50.57%**	-21.43%	-50.30%	-52.74%	-51.37%	-52.91%	-50.45%	-73.34%	-80.11%	-79.48%
	Upsampling	-40.17%	**+64.50%**	**+5.06%**	**+65.19%**	-3.40%	-33.38%	-38.38%	-35.24%	-38.36%	-34.70%	-60.00%	-60.08%	-60.06%
	PU learning	-16.45%	**+5.55%**	-1.97%	**+11.51%**	-1.75%	-3.81%	-12.23%	-6.79%	-13.28%	-8.69 %	-68.00%	-0.14%	-17.94%
	Relabeling	-34.33%	-19.47%	-3.51%	-42.46%	-3.17%	-28.30%	-32.37%	-29.61%	-32.51%	-31.30%	-62.36%	-20.02%	-28.31%
US-GB	Elasticsearch	-77.83%	-77.37%	-77.53%	-74.63%	-76.24%	-68.01%	-78.06%	-72.05%	-79.94%	-75.32%	-62.57%	**+112.45%**	**+1.77%**
	Pointwise	-75.04%	-60.04%	-66.70%	-68.49%	-71.57%	-76.76%	-74.42%	-72.40%	-75.80%	-71.39%	-97.20%	-98.84%	-98.70%
	Listwise	-77.83%	-77.37%	-77.53%	-74.63%	-76.24%	-68.01%	-78.06%	-72.05%	-79.94%	-75.32%	-37.52%	-75.00%	-70.96%
	Upsampling	-67.00%	**+20.30%**	-34.50%	**+11.55%**	-47.41%	-53.47%	-67.28%	-59.02%	-70.10%	-63.77%	-68.75%	-75.00%	-73.81%
	PU learning	-50.50%	-38.90%	-43.16%	-35.17%	-43.42%	-32.27%	-50.98%	-39.75%	-55.20%	-46.57%	-42.32%	-25.00%	-29.89%
	Relabeling	-33.19%	-19.65%	-24.50%	-28.28%	-30.68%	-44.34%	-33.85%	-43.45%	-40.74%	-31.43%	-50.00%	-75.00%	-71.72%

Table 2: Summarization of all baseline comparative performances. It reports all baseline normalized performance difference with X-SHOT model. Bold positive values (+) mean related baselines outperform X-SHOT model.

trieved skills shared in both locales, and %Common Labels is the rate of label counts on retrieved common skills. Testing utterances in both datasets tend to obtain less common skills than inference utterances.

Dataset	Type	%Comm. Skills	%Comm. Labels	#Label/Utt.	%Retrieval
US-CA	Testing	61.34%	48.56%	1.10	71.76%
	Inference	63.35%	69.24%	0.46	39.82%
US-GB	Testing	57.51%	44.16%	1.01	76.23%
	Inference	61.08%	71.06%	0.54	46.64%

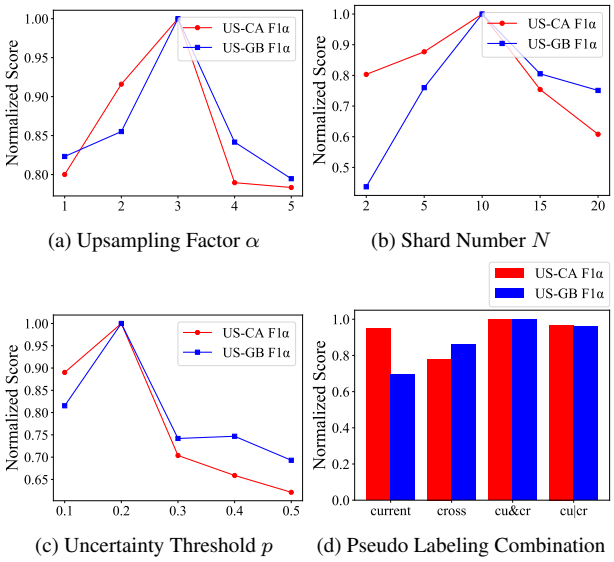
Table 3: X-SHOT model prediction result statistics.

3.6. Ablation studies

In Table 4, there are five components that can be disassembled from the full model, including three types of model inputs (Skill id, Skill name, and Skill score) and two types of model structures (Fusion layer and Bi-LSTM layer). We iteratively remove each component to report their normalized performance difference with the full model. In most cases, removing any component will lead to performances drop on all metrics. While in dataset US-CA, recall is slightly increased by removing skill name, score or fusion layer. But they sacrifice larger decreases on precisions and F1 scores. Moreover, removing skill id slightly increases Prescion$_a$ and F1$_a$ in dataset US-CA, but all its rest metrics decrease significantly. It means that although the retrieved skills are more precise, they are less generic and not located in the top ranked position of all retrieved skills. In dataset US-GB, all metrics drop when removing any component from the full model. We can conclude that each component has positive influence for cross-locale skill ranking. Bi-LSTM layer has the largest impact as removing it hurts model performance the most.

3.7. Parameter Tuning

There are four major parameters to tune, including biased upsampling factor α, shard number N, uncertainty threshold p and pseudo labeling combination. Figure 2 shows the F1$_\alpha$ score changes associated with different parameter setups. Due to the Alexa customer data confidential policy, we can only report the normalized scores calculated as the original scores divided by the largest metric score.



(a) Upsampling Factor α (b) Shard Number N

(c) Uncertainty Threshold p (d) Pseudo Labeling Combination

Fig. 2: Parameter tuning for X-SHOT model.

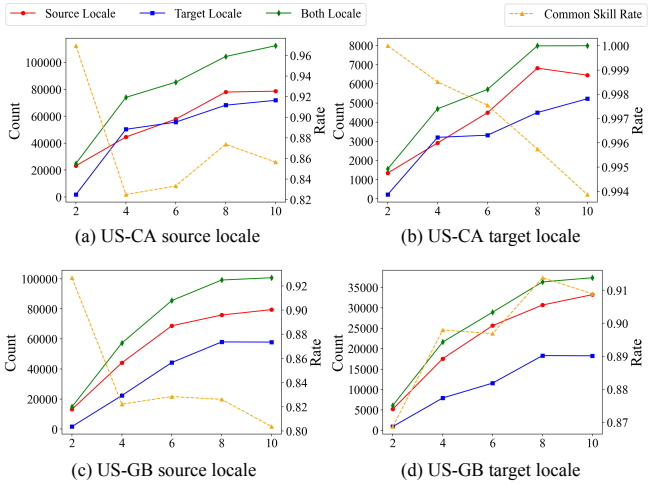
Figure 2(a) shows the upsampling factor α influence in two datasets. The best model is obtained when upsampling positive labeled data for three times. In Figure 2(b), if the training data is split into too many shards ($N > 10$), the model will not be well trained in each data shard to further predict adequate labels for next data shard. Figure 2(c) reveals $p = 0.2$ is the best setup to filter enough amount of high quality pseudo labels. Figure 2(d) shows four types of pseudo label generation, including from current locale labels (current), cross locale labels (cross), the interaction of two locale labels (cu&cr) and the union of two locale labels (cu|cr). It means adding intersection of pseudo labels generated from both locales returns the best result.

3.8. Dynamic Label Change

To demonstrate the dynamic changes of pseudo labels generated in each data shard, Figure 3 visualizes common skill rates (the rate of pseudo labels on common skills) and pseudo label count of two locales across data shards in two datasets. Under all circumstances, pseudo label count keeps increasing along with more training data shards, reflecting the growing

Dataset	Model	Classification Metrics					Ranking Metrics				
		Recall	Precision$_a$	F1$_a$	Precision$_b$	F1$_b$	Precision	Recall	F1	NDCG	MAP
US-CA	– Skill id	-22.64%	**+23.97%**	**5.49%**	-18.80%	-2.31%	-12.03%	-10.07%	-11.99%	-11.80%	-10.23%
	– Skill name	**+2.44%**	-30.87%	-23.63%	-29.89%	-19.67%	-41.89%	-1.31%	-31.43%	-6.77%	-18.46%
	– Skill score	**+5.26%**	-35.41%	-27.21%	-32.68%	-21.22%	-35.99%	-3.51%	-27.53%	-10.55%	-18.69%
	– Fusion layer	**+3.87%**	-44.21%	-35.53%	-39.16%	-26.96%	-46.05%	**+2.79%**	-33.60%	-6.82%	-20.18%
	– Bi-LSTM	-73.87%	-38.74%	-55.96%	-41.60%	-61.01%	-75.64%	-73.10%	-75.13%	-74.93%	-76.30%
US-GB	– Skill id	-39.67%	-20.52%	-27.79%	-25.40%	-32.88%	-49.21%	-53.62%	-51.42%	-59.09%	-57.83%
	– Skill name	-7.89%	-25.76%	-20.92%	-27.35%	-19.30%	-21.88%	-11.31%	-18.45%	-21.32%	-22.40%
	– Skill score	-3.56%	-14.97%	-11.69%	-14.73%	-9.80%	-5.32%	-5.57%	-5.40%	-14.93%	-16.62%
	– Fusion layer	-4.19%	-40.34%	-31.02%	-38.12%	-23.44%	-16.33%	-2.78%	-10.57%	-10.93%	-15.98%
	– Bi-LSTM	-84.67%	-62.79%	-74.37%	-64.46%	-78.09%	-82.54%	-84.83%	-83.57%	-86.80%	-86.11%

Table 4: Summarization for ablation studies. "-" refers to the removed model component and metric scores are the ablative model's normalized performance differences with the full model.



(a) US-CA source locale (b) US-CA target locale

(c) US-GB source locale (d) US-GB target locale

Fig. 3: Generated pseudo label counts as well as their common skill rates in each locale across data shards.

model prediction capability. However, the adaptive uncertainty threshold p restrains the growing speed to avoid generating noisy pseudo labels to pollute training data. As the union of two locale pseudo labels do not remarkably exceed individual locale pseudo labels, which reflects huge overlapping between the two locale generated pseudo labels. As for skill type distribution of pseudo labels, most of retrieved skills are common skills because of their high volumes in two datasets (in Table 3). In most of cases except Figure 3(d), the common skill rate drops during training process, revealing the learned Reranker becomes more and more capable to retrieve unique skills in its associated locale.

4. RELATED WORK

Skill ranking is a fundamental task in SLU systems, which aims to retrieve top relevant skills for utterances [16, 17]. [18] is a classic work which considers utterance contextual session information via a RNN unit to predict its relevant skills. [19] first applies knowledge distillation and PU learning techniques to augment pseudo labels for each utterance. [20] leverages skill classifier with a personalized attention module guided by customer enabled skills. [21] integrates skill retrieval, intent classification and slot filling into a unified model trained with shared encoders.

To extend the scope of skill candidates and improve model performance, cross-domain knowledge is utilized to bring in high-quality labelled data and external skills. [2, 3] propose multinomial adversarial networks to learn invariant features across multiple domains. [22] firstly learns a source domain classifier, then relies on a knowledge distillation approach to transfer the learned knowledge. Similarly, [23] maintains a mapping dictionary to train student model in target domain. [24] leverages an adversarial feature adaptation technique to distill discriminative knowledge across domains.

Positive-Unlabeled (PU) learning and co-training are two major types of self-training approaches for training data augmentation. [5] extends PU learning to also incorporate negative labeled data. [25] fine-tunes the learned model with reweighted pseudo labels via a teacher-student framework. [26] introduces a scalable PU learning approach and [27, 28] converts PU learning into a risk minimization problem. Co-training methods learn two classifiers on two views to describe the same instances, which jointly contribute to pseudo labeling on unlabelled data. [29] simultaneously trains multiple deep neural networks as different views and exploits adversarial examples to encourage view difference. [6] learns a teacher model on labelled data to guide cross-view training on unlabelled data. [30] designs an asymmetric co-teaching model which better resists noisy labels.

5. CONCLUSION

For skill ranking in small SLU locales with scarce developed skills and labeled utterances, we present a shard-based co-training method which exerts cross-locale knowledge to bring in new skills and pseudo labels for model enhancement. In the next step, we will put more efforts on improving model individual components, such as upgrading Shortlisting to neural network models or integrating Shortlisting with Reranking to achieve an end-to-end framework. Meanwhile, we will also explore more advanced co-training strategies to improve the quantity and quality of generated pseudo labels.

6. ACKNOWLEDGMENT

We thank Ming Tan, Beiye Liu, and Konstantine Arkoudas for their valuable opinions on this paper.

7. REFERENCES

[1] Zheng Gao, Radhika Arava, Qian Hu, Xibin Gao, Thahir Mohamed, Wei Xiao, and Mohamed AbdelHady, "Paraphrase label alignment for voice application retrieval in spoken language understanding," *Proc. Interspeech 2021*, pp. 4199–4203, 2021.

[2] Xilun Chen and Claire Cardie, "Multinomial adversarial networks for multi-domain text classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1226–1240.

[3] Yuan Wu and Yuhong Guo, "Dual adversarial co-learning for multi-domain text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6438–6445.

[4] Zheng Gao, Hongsong Li, Zhuoren Jiang, and Xiaozhong Liu, "Detecting user community in sparse domain via cross-graph pairwise learning," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 139–148.

[5] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama, "Classification from positive, unlabeled and biased negative data," in *International Conference on Machine Learning*, 2019, pp. 2820–2829.

[6] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le, "Semi-supervised sequence modeling with cross-view training," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1914–1925.

[7] Thomas Roelleke and Jun Wang, "Tf-idf uncovered: a study of theories and probabilities," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 435–442.

[8] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang, "Embedding-based retrieval in facebook search," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2553–2561.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, Aug. 2016, pp. 1715–1725, Association for Computational Linguistics.

[11] Clinton Gormley and Zachary Tong, *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*, " O'Reilly Media, Inc.", 2015.

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[13] Zhiheng Huang, Wei Xu, and Kai Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[14] Jessa Bekker and Jesse Davis, "Learning from positive and unlabeled data: a survey.," *Mach. Learn.*, vol. 109, no. 4, pp. 719–760, 2020.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019.

[16] Zheng Gao and Rui Bi, "University of pittsburgh at trec 2014 microblog track," Tech. Rep., PITTSBURGH UNIV PA SCHOOL OF INFO SCIENCES, 2014.

[17] Zheng Gao and John Wolohan, "Fast nlp-based pattern matching in real time tweet recommendation.," in *TREC*, 2017.

[18] Puyang Xu and Ruhi Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 136–140.

[19] Joo-Kyung Kim and Young-Bum Kim, "Pseudo labeling and negative feedback learning for large-scale multi-label domain classification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7964–7968.

[20] Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, and Ruhi Sarikaya, "Efficient large-scale neural domain classification with personalized attention," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2214–2224.

[21] Young-Bum Kim, Sungjin Lee, and Karl Stratos, "Onenet: Joint domain, intent, slot prediction for spoken language understanding," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 547–553.

[22] Yogarshi Vyas and Marine Carpuat, "Weakly supervised cross-lingual semantic relation classification via knowledge distillation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5288–5299.

[23] Giannis Karamanolakis, Daniel Hsu, and Luis Gravano, "Cross-lingual text classification with minimal resources by transferring a sparse teacher," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 3604–3622.

[24] Ruochen Xu and Yiming Yang, "Cross-lingual distillation for text classification," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 1415–1425.

[25] Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah, "Adaptive self-training for few-shot neural sequence labeling," *arXiv preprint arXiv:2010.03680*, 2020.

[26] E Sansone, FGB De Natale, and ZH Zhou, "Efficient training for positive unlabeled learning.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2584, 2019.

[27] Takashi Ishida, Gang Niu, and Masashi Sugiyama, "Binary classification from positive-confidence data," in *Advances in Neural Information Processing Systems*, 2018, pp. 5917–5928.

[28] Hong Shi, Shaojun Pan, Jian Yang, and Chen Gong, "Positive and unlabeled learning via loss decomposition and centroid estimation.," in *IJCAI*, 2018, pp. 2689–2695.

[29] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille, "Deep co-training for semi-supervised image recognition," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 135–152.

[30] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li, "Asymmetric co-teaching for unsupervised cross-domain person re-identification.," in *AAAI*, 2020, pp. 12597–12604.