X-SHOT: Learning to Rank Voice Applications via Cross-Locale Shard-based Co-Training

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

Amazon Alexa AI, Seattle, USA

{zhenggao, aravar, huqia, gxibin, thahirm, weixiaow, mbdeamz}@amazon.com

Abstract

Virtual assistants such as 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. 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. Extensive experimental results from two datasets demonstrate that our model significantly outperforms a number of strong alternatives.

Index Terms: spoken language understanding, cross-locale retrieval, co-training

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. For example, utterance "*alexa, play today's hits*" will directly invoke skill "*Pandora*" to play trending music in the device. There are over 100 million Alexa devices receiving billions of utterances everyday.

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, which limits their capability for utterance responses and ends up with a huge volume of unclaimed utterances having no suggested 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.

As locale can be regarded as a specific type of domain, cross-domain investigations can naturally solve target locale retrieval tasks by transferring source locale knowledge via adversarial training [1] or knowledge distillation [2] 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 learning [3] or self-training [4] techniques. However they only explore within-locale data and bring in no external knowledge.

Unclaimed utterances with no suggested skills will raise customer dissatisfaction, which 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, the selected source locale is intentionally with the same language as target locale. 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 two Reranker models trained from previous data shard will both make predictions on next data shard. Their predicted positive labels are jointly utilized as augmented labels to facilitate next shard training. In the testing stage, only source locale Reranker model is used to retrieve source locale skills to target locale unclaimed utterances.

2. Method

The proposed two-step listwise approach firstly retrieves the top K most relevant skills, then co-trains 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:

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

$U = \{U^s, U^t\}$ represents 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 $U_i = \{U_i^s, U_i^t\}$, we retrieve its top K most relevant skills from both locales via Shortlisting Elasticsearch indexes E^s and E^t [5]. After that, each utterance receives two skill lists. Further on, to alleviate unlabeled data influence, a data augmentation approach is applied on Shortlisting skills with a combination of biased upsampling and pseudo labeling. Then the augmented data is used to incrementally train Reranker models R^s and R^t where unknown/rejected skills are labeled as "negative". Figure 1 is a

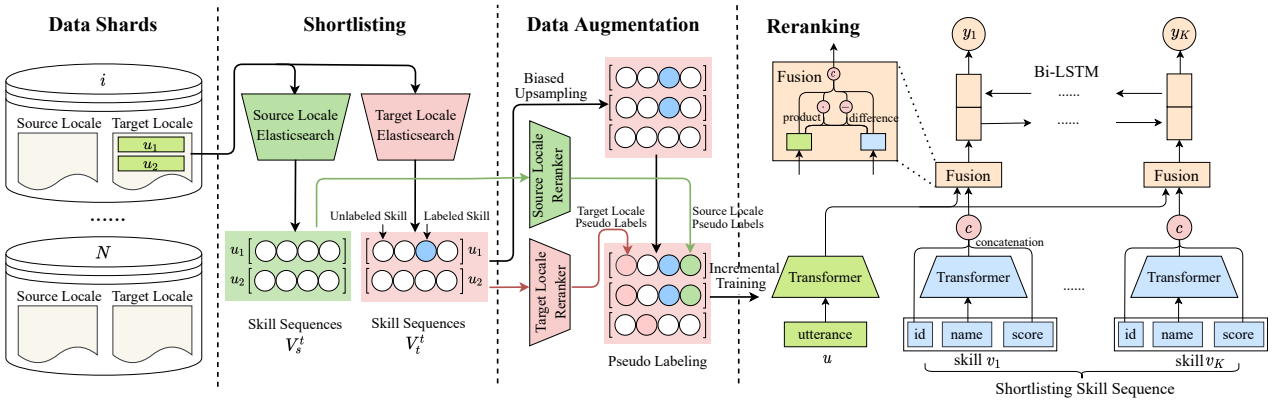

Figure 1: *X-SHOT model training in target locale. This figure shows how target locale Reranker model R^t is updated in the i_{th} data shard. The same training process happens in the source locale to train Reranker model R^s.*

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 without redundancy.

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;
    $Y_i^s \leftarrow Y_i^s + (L_s^s \cap L_s^s) \in V_s^s$;
    $Y_i^t \leftarrow Y_i^t + (L_s^t \cap L_t^t) \in V_t^t$;
  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

3. Experiments

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 English utterances collected from production in November 2020.

To evaluate the main contribution of this paper, six baselines are chosen from either skill ranking (Elasticsearch [5], Pointwise [6], Listwise [7]) or pseudo labeling (Upsampling positive instances, PU learning [8], Relabeling [9]) perspective. In the end, recall, precision and F1 score are reported.

Dataset	Model	Recall	Precision	F1
US-CA	Elasticsearch	-64.76%	-56.79%	-59.45%
	Pointwise	-70.81%	-19.55%	-46.43%
	Listwise	-53.99%	**+71.76%**	-4.34%
	Upsampling	-40.17%	**+64.50%**	**+5.06%**
	PU learning	-16.45%	**+5.55%**	-1.97%
	Relabeling	-34.33%	-19.47%	-3.51%
US-GB	Elasticsearch	-77.83%	-77.37%	-77.53%
	Pointwise	-75.04%	-60.04%	-66.70%
	Listwise	-77.83%	-77.37%	-77.53%
	Upsampling	-67.00%	**+20.30%**	-34.50%
	PU learning	-50.50%	-38.90%	-43.16%
	Relabeling	-33.19%	-19.65%	-24.50%

Table 1: *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.*

Table 1 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 the table, baselines perform similar trends in both datasets, while there are still unique patterns lying in each individual dataset. 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. Another finding is that Pointwise model performs worse than Listwise model, reflecting the convergence and optimization difficulty in Pointwise model and the superiority of two-step listwise approach.

4. 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 explore more advanced co-training strategies to improve the quantity and quality of generated pseudo labels to support model optimization.

5. References

[1] X. Chen and C. 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.

[2] Y. Wu and Y. Guo, "Dual adversarial co-learning for multi-domain text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6438–6445.

[3] Y.-G. Hsieh, G. Niu, and M. Sugiyama, "Classification from positive, unlabeled and biased negative data," in *International Conference on Machine Learning*, 2019, pp. 2820–2829.

[4] K. Clark, M.-T. Luong, C. D. Manning, and Q. 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.

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

[6] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[7] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

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

[9] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019.