Cross-lingual Transfer Learning with Data Selection for Large-Scale Spoken Language Understanding

Quynh Do
Amazon
Aachen, Germany
doquynh@amazon.com

Judith Gaspers
Amazon
Aachen, Germany
gaspers@amazon.de

Abstract

A typical cross-lingual transfer learning approach boosting model performance on a resource-poor language is to pre-train the model on *all* available supervised data from another resource-rich language. However, in large-scale systems, this leads to high training times and computational requirements. In addition, characteristic differences between the source and target languages raise a natural question of whether *source-language data selection* can improve the knowledge transfer. In this paper, we address this question and propose a simple but effective language model based source-language data selection method for cross-lingual transfer learning in large-scale spoken language understanding. The experimental results show that with data selection i) the source data amount and hence training speed is reduced significantly and ii) model performance is improved.

1 Introduction

Spoken Language Understanding (SLU) plays an important role in spoken language technology and is typically divided into two sub-tasks: Intent Classification (IC) and Slot Filling (SF). While the former identifies a speaker's intent, the latter extracts semantic constituents from the natural language query. Recently, there have been emerging efforts on Cross-Lingual Transfer Learning (CLTL) methods to reduce data requirements in deep neural network (DNN) based SLU. A typical approach is to pre-train the model on labeled data from a richly resourced language, and then either apply it directly on a target language (Upadhyay et al., 2018) or fine-tune it on a smaller amount of supervised data from a target language (Do and Gaspers, 2019). However, both in SLU and other NLP tasks, prior work on CLTL typically utilized all available source data to transfer knowledge, as it has been mainly investigated in rather small Academic settings. However, in large-scale settings with millions of utterances, this would lead to costly training times, high computational requirements and optimization difficulties. Moreover, the different characteristics between the source and target languages raise a natural question of whether source-language data selection in which only the most relevant source instances are picked for pre-training can improve CLTL performance, as a considerable amount of source utterances might be "irrelevant" to the target language or even yield negative transfer.

Addressing these questions, in this paper we explore source-language data selection for CLTL in SLU, focusing especially on large-scale settings in which we assume the existence of a *large* amount (millions) of source data and a *moderate* amount (thousands) of target data. Since the effectiveness of pre-training in CLTL depends on the similarity of the source data distribution and the real distribution of the target language, we propose a source-language data selection method which computes the relevance score of each source instance to the target language by using several N-gram language model based metrics. Our method is designed to satisfy *both* IC and SF sub-tasks in a multi-task training scenario, and to select data from *multiple* source languages, which have been rarely studied in the literature.

Our experimental results show that our proposed data selection method: i) improves CLTL performance in large-scale settings, while reducing the amount of source data significantly ii) brings higher gains to SF but does not hurt IC, and iii) can select data efficiently from multiple source languages for a single target language.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1455–1460,
Hong Kong, China, November 3–7, 2019.

2 Related work

Prior work on CLTL for SLU has mainly focused on using machine translation (e.g. García et al. (2012); He et al. (2013); Gaspers et al. (2018)). Until recently, few approaches based on cross-lingual joint training and cross-lingual supervised pre-training for DNNs have been proposed. The former takes advantage of the knowledge transferring in a SLU system by jointly training (relatively balanced) source and target data (e.g. Li et al. (2018)). Meanwhile, the latter is usually used when the amount of source data is significantly larger than the amount of target data. In particular, the SLU model is pre-trained on a large amount of supervised source data, and then either tested directly on the target language (e.g. Upadhyay et al. (2018)), or fine-tuned on a smaller amount of supervised target data (e.g. Do and Gaspers (2019)). Our CLTL method follows the line of Do and Gaspers (2019), but instead of utilizing all available source data, we aim at selecting the most relevant subset of the source data for the target language.

Data selection has been studied in the field of domain adaptation with most of the work targeting machine translation (Axelrod et al., 2011; van der Wees et al., 2017). These approaches usually rank sentence pairs in a large bitext from a source domain according to their difference in cross-entropy or perplexity with respect to a target domain corpus and then select the top n sentence pairs to train a machine translation system for the target domain. Although this task also deals with multiple languages, it is not a CLTL problem. The application of data selection on other tasks are relatively rare, e.g., dependency parsing (Plank and van Noord, 2011), sentiment analysis (Remus, 2012), POS tagging (Ruder and Plank, 2017).

Several common data metrics have been proposed to rank the relevance of the source instances to the target domain, e.g, word similarity measures, diversity. However, to the best of our knowledge, data selection has *not yet* been explored for DNN-based CLTL in SLU. In addition, two challenges tackled in this paper, i.e. applying data selection for a multi-task training scenario and dealing with multiple source languages, have been rarely studied in the literature.

3 Spoken language understanding

3.1 Task definition

Suppose for a language l with word vocabulary $\mathcal{V}_l$, intent vocabulary $\mathcal{I}_l$ and slot vocabulary $\mathcal{S}_l$, we have a set of utterances which are annotated with an intent label and each word is annotated with a slot label. The task of SLU is divided into two sub-tasks: i) *Intent classification*, which learns a function mapping each unlabeled utterance to a proper intent label $\in \mathcal{I}_l$, and ii) *Slot filling*, which learns a function mapping each unlabeled token to a proper slot label $\in \mathcal{S}_l$.

3.2 Model

Our multi-task SLU model consists of: i) A shared embedding layer which is the concatenation of a 1-dimensional convolution neural network based character embedding and a word embedding. ii) A shared encoder which is a two-layers bi-directional highway Long-short Term Memory network (Srivastava et al., 2015) served by the embedding layer as inputs, learning a contextual, fixed-dimensional representation for each token. ii) Two decoders for SF and IC sub-tasks; each consists of a stack of two dense layers and a softmax layer on top.

The two sub-tasks are trained jointly via a weighted loss function: $L = \alpha_i \hat{L}_i + \alpha_s \hat{L}_s$, where $\hat{L}_i, \hat{L}_s$ are the normalized cross-entropy losses with label smoothing (Szegedy et al., 2016) of IC and SF, respectively.

3.3 Cross-lingual transfer learning

Given a target language l^t with a limited supervised data set $\mathbf{D}_{l^t}$ divided into a training set $\mathbf{D}_{l^t}^T$ and a validation set $\mathbf{D}_{l^t}^V$, CLTL aims at improving the SLU performance on l^t by leveraging the larger supervised data sets $\mathbf{D}_{l_1^s}, \ldots, \mathbf{D}_{l_N^s}$ from N source languages $l_1^s, \ldots, l_N^s$. A common idea behind CLTL methods is to map the source and target data into a shared space, so that the knowledge can be transferred in-between languages.

In this paper, we assume the availability of a multi-lingual word embedding function which maps a word in any language into a shared space: $\mathcal{W} : \mathcal{V}_{l_1^s} \bigcup \ldots \mathcal{V}_{l_N^s} \bigcup \mathcal{V}_{l^t} \rightarrow \mathbb{R}^d$, where $\mathcal{V}_l$ is the vocabulary of language l. In addition, for a source language l_i^s in which a bilingual dictionary $\mathcal{D}_{l_i^s, l^t} : \mathcal{V}_{l_i^s} \rightarrow \mathcal{V}_{l^t}$ is available, a word w can be alternatively mapped into the shared space by using $\mathcal{W}(\mathcal{D}_{l_i^s, l^t}(w))$ [1]. The word embedding layer in our

[1] Experiments on the development data shows that

SLU model is fixed to the mapping from the word vocabulary to this shared space, without being updated during training. In contrast, the character embedding layer in our model is initialized randomly and updated during training. Our CLTL training strategy consists of two phases: First, the model is *pre-trained* on the source data $\mathbf{D}_{l_1^s} \cup \ldots \cup \mathbf{D}_{l_N^s}$ for T_w^s epochs, and validated on $\mathbf{D}_{l^t}^V$. Second, the model is *fine-tuned* on the target data $\mathbf{D}_{l^t}^T$ for T_w^t epochs and validated on $\mathbf{D}_{l^t}^V$.

4 Data selection

The effectiveness of pre-training in CLTL depends on the similarity of the source data distribution and the real distribution of the target language. Let us consider each component in our model. First, for the word embedding layer, obtaining similar distributions of the source data and the target language can be considered as an "easy" task by using multilingual word embedding. Second, for the character embedding layer and the encoder, it depends on how similar the character patterns and the word patterns of the source data and the target language are, respectively. Finally, for the decoders, the similar distributions could be expected given the good distributions provided by the encoder.

We, therefore, propose a relevance metric for the source utterances w.r.t. the target language: $R(u) = \sum_{k=1}^{M} \alpha_k f_k(u)$, where f_k and α_k are respectively an attribute value and its importance weight, and M is the total number of attributes. Each attribute is associated with an N-gram word- or character- based language model trained on the target language which can be used to estimate the similarity of a pattern to the target language. Let us consider an attribute f_k and its N-gram language model LM_k trained on the target language l_t. Given an utterance $u = w_1 \ldots w_n$ in a source language l_i^s and the bilingual dictionary $\mathcal{D}_{l_i^s, l^t} : \mathcal{V}_{l_i^s} \rightarrow \mathcal{V}_{l^t}$ mapping a word in l_i^s to another word in l^t, we call $\mathbf{S}$ the set of N-grams generated from $\mathcal{D}_{l_i^s, l^t}(w_1) \ldots \mathcal{D}_{l_i^s, l^t}(w_n)$. The attribute value f_k is computed as the average language model score of the elements in $\mathbf{S}$:

$$f_k(u) = \sum_{g \in \mathbf{S}} \mathrm{LM}_k(g) * \frac{1}{|\mathbf{S}|} \qquad (1)$$

$\mathcal{W}(\mathcal{D}(w))$ works well for French as target language, while $\mathcal{W}(w)$ is better for German.

Exp.	Source	Target		
		Train	Dev	Test
10K-DE	EN 5M	10K	2K	7,431
20K-DE	EN 5M	20K	2K	7,431
10K-DE	EN 5M, DE 1.1M, ES 114,702	10K	5K	58K
20K-DE	EN 5M, DE 1.1M, ES 114,702	20K	5K	58K

Table 1: Supervised data statistics.

We then normalize $f_k(u)$ at intent level:

$$\bar{f}_k(u) = \frac{f_k(u)}{\max_{u' \in \mathbf{I}_u} f_k(u')} \qquad (2)$$

where $\mathbf{I}_u$ is the set of utterances (from all source languages) having the same intent as u. By using the proposed relevance metric, the source data can be ranked in descending order, and only the top-K utterances will be selected for the pre-training.

5 Experiments

5.1 Data

Supervised data For large-scale experiments, we extracted random samples from a large-scale SLU system. The data are representative of user requests to voice-enabled devices and are labeled with intents and slots. We include four languages into our experiments, i.e. English (EN), German (DE), French (FR) and Spanish (ES). DE and FR are used as the target languages in our experiments. Data statistics can be found in Table 1.

Unlabelled data sets For each of the target languages (DE and FR), we build N-gram language models on unlabelled data sets in that language. We make use of 3M DE sentences and 1M FR sentences which are freely available from the Leipzig unlabelled corpus collection (Goldhahn et al., 2012). In addition, we collect 500K DE and 2.5K FR unlabelled utterances with a similar nature as labelled utterances from the SLU system.

Pre-trained resources We use pre-trained 300-dimensional multilingual word embeddings and bilingual dictionaries provided by Conneau et al. (2017) and Lample et al. (2017), respectively.

5.2 Setup

We carry out four experiments with 10K and 20K target data in DE and FR (see Table. 1 for the experiment names and the labelled data statistics). We conduct experiments with transferring from one source language (EN) to another (DE) and from

three source languages (EN, DE, ES) to one target language (FR). In each experiment, we compare four different training strategies: i) Mono (w.o. CLTL): an SLU model is trained on *only* the supervised *target* data. ii) CLTL: *all* supervised source data is used for pre-training, and we fine-tune on target language data. iii) CLTL-RD: random K% utterances of the original source data are used for pre-training. iv) CLTL-DS: the K% most relevant source utterances are selected by using our proposed relevance metric for pre-training.

Settings The convolutions used for character embeddings have window sizes of 2, 3, 4, each consisting of 64 filters. The sizes of LSTM and dense layers are set to 300. All dropout keep probabilities are set to 0.9. The hyper-parameter tuning on the development set results in $\alpha_i = 0.2, \alpha_s = 0.8$, label smoothing rate $= 0.1$. We use Adam optimizer with learning rate $= 0.001$. Exponential decay is applied to the learning rate with decay steps $= 500$, and decay rate $= 0.95$. For CLTL, the number of training epochs T_w^s and T_w^t are set to 6 and 25, respectively. For data selection, we use four N-gram language models, i.e. word-based bi-gram and tri-gram language models and character-based bi-gram and tri-gram language models. The four importance weights are set to 1.0 each. For evaluation we use the standard metrics, i.e. F1, precision and recall for slot filling (computed using the CoNLL 2002 script) and accuracy for intent classification.

5.3 Results and discussions

Exp.	Model	Slot			Intent
		P	*R*	*F1*	*Acc.*
10K-DE	Mono	76.4 ± 1.6	75.4 ± 1.4	75.9 ± 1.5	87.9 ± 0.4
	CLTL	79.7 ± 1.8	77.6 ± 1.1	78.7 ± 1.5	89.5 ± 0.3
	CLTL-RD	79.3 ± 0.9	77.0 ± 0.5	78.1 ± 0.6	89.5 ± 0.4
	CLTL-DS	80.1 ± 0.6	78.6 ± 0.7	**79.4 ± 0.6**	**90.0 ± 0.3**
20K-DE	Mono	80 ± 0.2	78.8 ± 1.0	79.4 ± 0.6	89.5 ± 0.0
	CLTL	81.3 ± 1.8	78.9 ± 2.3	80.1 ± 2.1	90.5 ± 0.1
	CLTL-RD	80.7 ± 1.6	79.3± 0.9	80.0 ± 1.2	90.1 ± 0.3
	CLTL-DS	82.2 ± 0.2	80.7 ± 0.1	**81.5 ± 0.1**	**90.8 ± 0.4**
10K-FR	Mono	76.5 ± 0.5	78.4 ± 0.5	77.5 ± 0.5	89.3 ± 0.4
	CLTL	79.0 ± 0.3	80.7 ± 0.4	79.8 ± 0.1	90.7 ± 0.1
	CLTL-RD	78.7 ± 0.2	80.6 ± 0.3	79.7 ± 0.2	90.4 ± 0.3
	CLTL-DS	80.0 ± 0.7	82.0 ± 0.2	**81.0 ± 0.3**	**91.2 ± 0.2**
20K-FR	Mono	78.9 ± 0.1	80.2 ± 0.2	79.5 ± 0.2	90.6 ± 0.1
	CLTL	80.4 ± 0.6	82.4 ±0.5	81.4 ± 0.6	91.4 ± 0.2
	CLTL-RD	80.9 ± 0.1	82.5 ± 0.5	81.7 ± 0.2	91.5 ± 0.1
	CLTL-DS	81.5 ± 0.2	82.8 ± 0.2	**82.1 ± 0.2**	**91.6 ± 0.2**

Table 2: Performance of CLTL on large-scale data sets. K is set to 50 (%) in CLTL-RD and CLTL-DS. Reported results are the mean and std. values of 3 runs.

Is 100% better than 50%? Table 2 shows the performances of the different training strategies in our experiments. In CLTL-RD and CLTL-DS settings, K is set to 50 (%). It helps to answer the question that whether using the full source data (100%) is better than using just 50%. Interestingly, although 100% (CLTL) is better than random 50% (CLTL-RD) in 3 out of 4 experiments, it is surpassed by our selected 50% in all of the experiments. These results do not only prove the effectiveness of our proposed data selection metric but also suggest a potentially powerful application of source-language data selection on cross-lingual transfer learning.

Slot filling vs. intent classification As shown in Table 2source-language, our data selection method tends to bring higher gains to SF than to IC. One possible reason is that IC is the easier between the two sub-tasks (less categories, single label vs. sequence label decoding etc.). However, it is important to stress that our data selection does not hurt IC, meaning that the method is useful for joint learning.

One vs several source languages The similar trends in experiments with DE as target and FR as target show that our proposed data selection method can be applied in both single-source and multi-source transfer learning. In order to compare the utility of multiple vs a single source, we ran an experiment on 10K-FR using only English as the source language. The means of slot F1 and IC dropped from 81.0% and 91.2% to 79.9% and 90.8%, respectively, when using only English, potentially because it is not the closest language to French. Hence, the model could probably choose better source utterances from multiple sources. This indicates that our method works for both settings with higher gains for using multiple source languages.

Value of K and importance weights of the language models? One may question whether it is possible to optimize the importance weights of the language models and choose K automatically. A possible solution could be using Bayesian optimization (Ruder and Plank, 2017). Instead of selecting the top K% utterances, we define a threshold θ: u is selected if $R(u) >= \theta$. θ and α_k become hyper-parameters which can be optimized using Bayesian Optimization with the score of the CLTL model on a development set as the objective. However, while

this could improve performance, it is expensive, especially in large-scale settings, i.e. optimization would likely add more computation time than what we can gain by training on a subset.

Performance on a small-scale benchmark dataset While we are mainly interested in large-scale SLU, we also strive for further understanding of CLTL by conducting a similar experiment on a small-scale widely-used SLU benchmark data set, i.e. ATIS (Tür et al., 2010). It contains audio recordings and corresponding annotated transcriptions in English of people making flight reservations.

To compare with the state-of-the-art systems, we apply our monolingual model on ATIS. The model reaches 95.6% F1 for slot filling and 96.8% accuracy for intent classification which are *comparable* to the state-of-the-art results reported on the same data set (Do and Gaspers, 2019).

We then perform a cross-lingual transfer learning experiment from English to German on ATIS. To construct the training sets of the target language, we select two random subsets of 200 and 400 English utterances from the training part of ATIS and translate them into German. The development set of the target language is formed by the German translation of a random subset of 144 English utterances selected from the validation part of ATIS. The test set of the target language is the German translation version of the ATIS test set which includes 893 utterances. The annotated source data comprise 4015 English training utterances from ATIS.

Exp.	Model	Slot			Intent
		P	*R*	*F1*	*Acc.*
200-DE	Mono	79.8 ± 0.3	80.6 ± 0.6	80.2 ± 0.6	85.8 ± 0.6
	CLTL	81.8 ± 0.5	82.8 ± 0.9	82.3 ± 0.6	85.3 ± 1.6
	CLTL-RD	83.0 ± 1.6	83.3 ± 1.7	83.1 ± 1.6	86.8 ± 1.4
	CLTL-DS	84.8 ± 2.7	85.3 ± 2.3	**85.0 ± 2.5**	**87.2 ± 1.9**
400-DE	Mono	86.8 ± 1.1	87.7 ± 0.6	87.2 ± 0.8	88.2 ± 0.7
	CLTL	88.0 ± 0.5	87.8 ± 0.7	87.9 ± 0.6	88.0 ± 0.3
	CLTL-RD	87.8 ± 1.3	87.4 ± 1.1	87.6 ± 1.1	87.9 ± 0.8
	CLTL-DS	88.3 ± 0.1	88.0 ± 0.3	**88.2 ± 0.2**	**88.2 ± 0.2**

Table 3: Performance of CLTL on ATIS. K is set to 50 (%) in CLTL-RD and CLTL-DS. Reported results are the mean and std. values of 3 runs.

Table 3 shows the performance of CLTL on the small-scale ATIS data set. As can been seen, in general, CLTL with data selection (CLTL-DS) is still beneficial in a small-scale setting.

6 Conclusions

We presented an efficient approach to select source data for cross-lingual transfer learning in large-scale SLU. Our results indicate that by using data selection we can both improve performance and reduce source data significantly without a negative effect on system performance, which can reduce training time and computational requirements in large-scale systems greatly. This suggests an interesting future research direction toward data selection for cross-lingual transfer learning problems.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Quynh Ngoc Thi Do and Judith Gaspers. 2019. Cross-lingual transfer learning for spoken language understanding. *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Fernando García, Lluís F. Hurtado, Encarna Segarra, Emilio Sanchis, and Giuseppe Riccardi. 2012. Combining multiple translation systems for spoken language understanding portability. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*, pages 194–198.

Judith Gaspers, Penny Karanasou, and Rajen Chatterjee. 2018. Selecting machine-translated data for quick bootstrapping of a natural language understanding system. *Proceedings of NAACL-HLT*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, pages 759–765. European Language Resources Association (ELRA).

X. He, L. Deng, D. Hakkani-Tur, and G. Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8342–8346.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Yujiang Li, Xuemin Zhao, Weiqun Xu, and Yonghong Yan. 2018. Cross-lingual multi-task neural architecture for spoken language understanding. In *Proc. Interspeech 2018*, pages 566–570.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1566–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. Remus. 2012. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 717–723.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382. Association for Computational Linguistics.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2377–2385. Curran Associates, Inc.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2010. What is left to be understood in atis? In *SLT*, pages 19–24. IEEE.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Z. Hakkani-Tür, and Larry P. Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 6034–6038.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. *CoRR*, abs/1708.00712.