TRANSFER LEARNING OF LANGUAGE-INDEPENDENT END-TO-END ASR WITH LANGUAGE MODEL FUSION

Hirofumi Inaguma¹, Jaejin Cho², Murali Karthick Baskar³, Tatsuya Kawahara¹, Shinji Watanabe²

¹Graduate School of Informatics, Kyoto University, Kyoto, Japan
²Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
³Brno University of Technology Speech@FIT and IT4I Center of Excellence Brno, Czech Republic

ABSTRACT

This work explores better adaptation methods to low-resource languages using an external automatic speech recognition (ASR) system in a unified sequence-to-sequence (S2S) architecture. We first build a language-independent ASR system using an external language model (LM) under the framework of transfer learning. When an external LM is integrated into the decoder network of the attention-based S2S model in the whole adaptation stage, to effectively incorporate linguistic context of the target language. We also investigate various seed models for transfer learning. Experimental evaluations using the IARPA BABEL data set show that LM fusion transfer improves performances on all target five languages compared with simple transfer learning when the external text data is available. Our final system drastically reduces the performance gap from the hybrid systems.

Index Terms— end-to-end ASR, multilingual speech recognition, low-resource language, transfer learning

1. INTRODUCTION

Fast system development for low-resourced new languages is one of the challenges in automatic speech recognition (ASR). Recently, end-to-end ASR systems based on the sequence-to-sequence (S2S) architecture [1,2] are filling up the gap of performance from the conventional HMM-based hybrid systems and showing promising results in many tasks with its extremely simplified training and decoding schemes [3,5]. This is very attractive when building systems in new languages quickly. However, models tend to suffer from the data sparseness problems in the low-resource scenario, especially in S2S models due to its data-driven optimization.

One possible approach to this problem is to utilize data of other languages. There are various approaches to leverage other languages: (a) to train a model multilingually (multi-task learning with other languages), and then further fine-tune to a particular language [6], and (b) to adapt a multilingual model to a new language using transfer learning [6,9] and additional features obtained from the multilingual model such as multilingual bottleneck features (BNF) [10,13] and language feature vectors (LFV) [14] (cross-lingual adaptation). To obtain a multilingual S2S model, a part of parameters can be shared while preparing the output layers per language [6], and we can further use a unified architecture with a shared vocabulary among multiple languages [15,17]. Since it would take much time to train such systems from scratch for many languages including new languages, we focus on the cross-lingual adaptation approach (b).

While a majority of the conventional transfer learning is concerned with acoustic model, using linguistic context during adaptation has not been investigated yet. The research question in this paper is: Is linguistic context also helpful for adaptation to new languages? The most common approach to integrate the external LM is referred to as shallow fusion, where LM scores are interpolated with scores from the S2S model [5,18,19]. Recently, several methods to leverage an external LM during training of S2S models are proposed: deep fusion [20] and cold fusion [21]. In deep fusion, the decoder network in the pre-trained S2S model and an external RNNLM are integrated into a single architecture by the gating mechanism and only the gating part is re-trained. In contrast, cold fusion integrates an external LM during the entire training stage.

In this paper, we investigate methods to fully utilize text data for adaptation to unseen low-resource languages. We propose LM fusion transfer, where an external LM is integrated into the decoder network of the S2S model only in the adaptation stage, as an extension of cold fusion. Since the decoder network is already well-trained in a language-independent manner, the model can better incorporate linguistic context from the external LM. The extra cost to integrate the external LM during adaptation is trivial in the resource constrained condition. We also investigate various seed multilingual models trained with 600 to 2200-hours speech data and show the effect of the amount and variety of multilingual training data.

Experimental evaluations on the IARPA BABEL corpus show that the LM fusion transfer improves performance compared to simple transfer learning with shallow fusion when the additional text data is available. The performance of the transferred models is drastically improved by increasing the model capacity and incorporating the external LM, and the resulting models perform comparably with the latest BLSTM-HMM hybrid systems [10]. To our best knowledge, this is the first results for the S2S model to show the competitive performance to the conventional hybrid systems in the low-resource scenario (~50 hours).

2. RELATED WORK

The traditional usage of unpaired text data in the S2S framework is categorized to four approaches: LM integration, pre-training, multi-task learning (MTL), and data augmentation. In the LM integration approach, there are three methods: shallow fusion, deep fusion, and cold fusion as described in Section 1. Their differences are in the timing to integrate an external LM and the existence of additional parameters of the gating mechanism. We depict these fusion methods in Fig. 1 [11]. In [19], these fusion methods are compared in middle-size English conversational speech (~300h) and large-scale Google

1 Although we can perform LM fusion during training of the seed multilingual model, we focus on applying it during adaptation because our goal is to adapt it to a particular language rapidly.
voice search data. However, none of previous works investigated the effect of them in other languages especially for low-resource languages, which is the focus of this paper. In [21], the authors show the effectiveness of cold fusion in a cross-domain scenario. Since the external LM is more likely to be changed frequently than the acoustic model, it is time-consuming to train a new S2S model with the LM integration from scratch every time the external LM is updated. In this work, we conduct LM fusion during adaptation, which is regarded as a more realistic scenario.

Another usage of the external LM is to initialize the lower layer in the decoder network with the pre-trained LM [19][23]. However, we transfer almost all parameters in a multilingual S2S model (including the decoder network), and thus we do not explore this direction. Apart from the external LM, the MTL approach with LM objective are investigated in [19][23]. Although the MTL approach does not require any additional parameters, it gets minor gains compared to LM fusion methods [19].

Recently, data augmentation of speech data based on text-to-speech (TTS) synthesis is investigated in the S2S framework [24][25]. Since we are interested in the usage of linguistic context during adaptation, we leave this direction to the future work.

3. END-TO-END ASR

3.1. Attention-based sequence-to-sequence

We build all models with attention-based sequence-to-sequence (S2S) models, which can learn soft alignments between input and output sequences of variable lengths [11]. They are composed of encoder and decoder networks. The encoder network transforms input features $x = (x_1, \ldots, x_T)$ to a high-level continuous representation $h = (h_1, \ldots, h_T') (T' \leq T)$, interleaved with subsampling layers to reduce the computational complexity [26]. The decoder network generates a probability distribution $P_{S2S}$ of the corresponding $U$-length transcription $y = (y_1, \ldots, y_U)$ conditioned over all previous generated tokens:

$$s_{u:S2S} = \text{Decoder}(s_{u-1:S2S}, y_{u-1}, c_u)$$

$$P_{S2S}(y|x) = \text{softmax}(W_{\text{out}} s_{u:S2S} + b_{\text{out}})$$

where $W_{\text{out}}$ and $b_{\text{out}}$ are trainable parameters, $s_{u:S2S}$ is a decoder state at the $u$-th timestep, and $c_u$ is a context vector summarizing notatable parts from the encoder states $h$. We adopt the location-based scoring function [2]. To encourage monotonic alignments, the auxiliary Connectionist Temporal Classification (CTC) [27] objective is linearly interpolated [28].

During the inference stage, hypotheses are generated with a beam search in a left-to-right manner. To avoid generating incomplete and repeated hypotheses, scores from the softmax layer used for the CTC objective are also taken into account [4]. Since CTC is frame-synchronous, it can detect the end point more accurately.

3.2. LM fusion

3.2.1. Shallow fusion

In the conventional decoding paradigm with an external LM, referred to as shallow fusion, scores from both the S2S model and LM are linearly interpolated to maximize the following criterion:

$$y^* = \arg \max_{y \in \mathbb{Y}^*} \{ \ln P_{\text{ASR}}(y|x) + \beta \ln P_{\text{LM}}(y) \}$$

$$s_{u:LM} = W_{LM} d_{u:LM} + b_{LM}$$

$$s_{u:S2S} = \sigma(W_{S2S} [s_{u:S2S}; s_{u:LM}] + b_{S2S})$$

$$s_{u:CF} = W_{CF} [s_{u:S2S}; s_{u:LM}] + b_{CF}$$

$$P_{S2S}(y|x) = \text{softmax}(\text{ReLU}(W_{\text{out}} s_{u:CF} + b_{\text{out}}))$$

where $W^{*}$ and $b^{*}$ are trainable parameters, $d_{u:LM}$ is a hidden state of RNNLM, $s_{u:LM}$ is a feature from the external LM, $s_{u:S2S}$ is a bottleneck feature before the final softmax layer, $g_u$ is a gating function, and $\odot$ represents element-wise multiplication. ReLU non-linear function is inserted before the softmax layer as suggested in [21]. We use the hidden state as a feature from RNNLM instead of logits because we use the universal character vocabulary for multilingual experiments, which results in the large softmax layer and increases the computational time [19].

In the original formulation in [21], scores from the external LM are not used. We found that linear interpolation of log probabilities from the LM with those from the S2S model during the inference as in shallow fusion still has complementary effects to improve performance. Therefore, we adopt it in all experiments.

3.2.2. Deep fusion (fine-tuning fusion)

Deep fusion [20] is another method to integrate an external LM during training. Unlike cold fusion, deep fusion is applied in the fine-tuning stage after both the S2S model and RNNLM are trained and fixed. Although deep fusion is formulated with a scalar gating function in [20], we use the same architecture as cold fusion in Section 3.2.2 to make a strict comparison. Then, the difference from the cold fusion is in the timing to integrate the external LM (see Figure 1). Note that the CTC loss is not used in the fine-tuning stage.
4. TRANSFER LEARNING OF MULTILINGUAL ASR

4.1. Adaptation to a target language

We adapt a seed language-independent ASR model to an (unseen) target language. We investigate the following four scenarios:

- **multi10**: From non-target 10 languages to an unseen target language
- **high2**: From 2 high-resource languages (English and Japanese) to an unseen target language
- **multi10+high2**: From the mix of non-target 10 languages and 2 high-resource languages to an unseen target language
- **multi15**: From the mix of non-target 10 languages and target 5 languages to a particular target language

The top three conditions are regarded as cross-lingual adaptation and the last one as multi-task learning with other languages followed by fine-tuning.

4.2. LM fusion transfer

During adaptation, all parameters are copied from the seed multilingual S2S model, then training is continued toward a target language. We investigate improved adaptation methods by integrating the external LM during and/or after transfer learning from the seed model. Three methods are considered as follows:

- **Transfer + SF**: Shallow fusion in Section 3.2.1 is conducted in the inference stage after adaptation.
- **Cold fusion transfer (CF-transfer)**: Cold fusion in Section 3.2.2 is conducted during adaptation. We integrate the external RNNLM from the start point of adaptation to a target language. The softmax layer is randomly initialized when adding the gating function.
- **Deep fusion transfer (DF-transfer)**: Deep fusion in Section 3.2.3 is conducted as fine-tuning after transfer learning. The softmax layer is randomly initialized when adding the gating function.

5. EXPERIMENTAL EVALUATION

5.1. Experimental setting

We used data from the IARPA BABEL project [29] and selected 10 languages as non-target languages for training the seed multilingual model: Cantonese, Bengali, Pashto, Turkish, Vietnamese, Haitian, Tamil, Kurmanji, Tokpisin and Georgian. For both monolingual and multilingual experiments, we used the same architecture. The softmax layer is randomly initialized when adding the gating function.

5.2. Results

### 5.2.1. Baseline monolingual systems for target 5 languages

First, we present the results of the baseline monolingual end-to-end systems in Table 1. Our new systems (line 2) significantly outperformed the old baseline reported on [7]. The gain mostly came from adding VGG blocks before BLSTM encoder and one more decoder LSTM layer though we also tuned other hyper-parameters. Next, changing the unit sizes of each LSTM layer from 320 to 1024 drastically improved the performance. This is surprising because increasing the number of parameters often makes the model overfit to the small amount of training data. Finally, shallow fusion with the monolingual RNNLM further boosted the performance although the RNNLM was trained with the small amount of parallel data only. We use this setting as default in the rest of experiments.

| Model | AS (54h) | SW (39h) | LA (58h) | TA (75h) | ZU (54h) |
|-------|---------|--------|--------|--------|--------|
| Old baseline | 73.9 | 66.5 | 64.5 | 73.6 | 76.4 |
| New baseline | 64.5 | 56.6 | 56.2 | 56.4 | 69.5 |
| + large units | 59.9 | 50.9 | 51.7 | 52.7 | 65.5 |
| + shallow fusion | 57.4 | 46.5 | 49.8 | 49.9 | 62.9 |
| BLSTM-HMM | 49.1 | 38.3 | N/A | 46.3 | N/A |

### 5.3. Results of cross-lingual adaptation

| Model | AS (54h) | SW (39h) | LA (58h) | TA (75h) | ZU (54h) |
|-------|---------|--------|--------|--------|--------|
| Cold fusion transfer (CF-transfer) | 78.1 | 62.6 | 62.1 | 71.6 | 78.4 |
| Deep fusion transfer (DF-transfer) | 75.9 | 57.6 | 57.1 | 57.4 | 66.5 |

### 5.4. Results of multi-task learning

| Model | AS (54h) | SW (39h) | LA (58h) | TA (75h) | ZU (54h) |
|-------|---------|--------|--------|--------|--------|
| Cold fusion transfer (CF-transfer) | 78.1 | 62.6 | 62.1 | 71.6 | 78.4 |
| Deep fusion transfer (DF-transfer) | 75.9 | 57.6 | 57.1 | 57.4 | 66.5 |

Table 1: Results of baseline monolingual systems. None of adaptation methods is conducted.

---

Increasing the unit size did not lead to the improvement.

---
end-to-end systems with shallow fusion are close to those of the hybrid system, just 3.6% absolute difference for Tagalog.

5.2.2. Comparison of seed language-independent models

We compared the seed language-independent models for adaptation to target languages. All models were transfered, and shallow fusion with the corresponding monolingual RNNLM trained with the parallel data was performed. The results are shown in Table 2. The overall performance was significantly improved by transfer learning. For Tagalog, the transfered S2S models achieved comparable WER to BLSTM-HMM in Table 1. We can see that multi10 model is generally better than high2 model despite the smaller data size, and combination of them (multi10+high2) gives slight improvement. On the other hand, multi15 model that includes the target language does not lead to further improvement. We can conclude that the diversity of languages is more important than the total amount of training data, and 10 languages are almost sufficient for learning language-independent feature representation and generalized to other languages well. In the following experiments, we use multi10 as the seed model and investigate cross-lingual adaptation.

5.2.3. Effect of LM fusion transfer

The results of our proposed LM fusion transfer are given in Table 3. When training S2S models from scratch, there is no difference among all fusion methods. When transfered from the multilingual S2S model, significant improvement is observed by integrating the external RNNLM. Shallow fusion was more effective than when training the S2S models from scratch in Table 1 because the multilingual training led to generalization and the affinity for the external LM was enhanced. CF-transfer got some improvements compared to transfer learning with shallow fusion for 3 target languages, but the effects of DF-transfer and CF-transfer are not significant. This is because RNNLMs were trained with only small parallel data, therefore linguistic context during adaptation was not so effective. However, CF-transfer in Tagalog outperformed the monolingual hybrid system in Table 1. When compared to the previous work using the same data, CF-transfer yielded 21.6% gains relatively on average. Furthermore, 6.8% gains were achieved from transfer learning without the external LM.

To investigate the effect of additional text data, we evaluate the LM fusion transfer with LLP on each target language (~10 hours). The results are shown in Table 4. We used monolingual RNNLM trained with LLP and FLP (~50 hours), respectively. The latter setting of a small speech data set (~10 hours) and a larger text data set (~50 hours) is regarded as a more realistic scenario in low-resource languages. When training S2S models from scratch, all models could not converge in our implementation even when reducing the unit sizes from 1024 to 320. The Babel corpus is mostly composed of conversational telephone speech (CTS), so it is difficult to optimize the S2S model from scratch with just around 10-hour training data. In the transfer learning approach, all three fusion methods got significant gains by using the external LM except for deep fusion in Assamese. For RNNLM trained with LLP, all fusion methods achieved a larger improvement than in Table 3, but similar performance. Interestingly, WER significantly dropped even when each RNNLM was trained with 10-hour data only. But all fusion methods show similar performance. In contrast, CF-transfer significantly outperformed simple transfer learning with shallow fusion on all 5 target languages when the RNNLM was trained with FLP, which is five-times larger than LLP. Therefore, we can conclude that linguistic context is helpful for adaptation when additional text data is available. This shows CF-transfer in Table 3 has the potential to surpass transfer learning with shallow fusion if we can access to additional text data. In summary, CF-transfer yielded relative 10.4% and 2.3% gains on average compared to transfer learning without and with shallow fusion, respectively.

6. CONCLUSION

We explored the usage of linguistic context from the external LM during adaptation of the language-independent S2S model to target low-resource languages. We empirically compared various LM fusion methods and confirmed its effectiveness in resource limited situations. We showed that cold fusion transfer is more effective than simply applying shallow fusion after adaptation when additional text is available, which means linguistic context is also helpful in addition to acoustic adaptation. Our S2S model drastically closed the gap from the BLSTM-HMM hybrid system.

Since the provided data only can be used for system training in BABEL rules, we do not explore to crawl text data from the WEB.
7. REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” in Proc. of ICLR, 2015.

[2] Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, “Attention-based models for speech recognition,” in Proc. of NIPS, 2015, pp. 577–585.

[3] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhiheng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Katyona Gonina, et al., “State-of-the-art speech recognition with sequence-to-sequence models,” in Proc. of ICASSP, 2018, pp. 4774–4778.

[4] Shinya Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

[5] Albert Zeyer, Kazuki Irie, Ralf Schluter, and Hermann Ney, “Improved training of end-to-end attention models for speech recognition,” in Proc. of Interspeech, 2018, pp. 7–11.

[6] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black, “Sequence-based multi-lingual low resource speech recognition,” in Proc. of ICASSP, 2018, pp. 4909–4913.

[7] Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiessner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiát, Shinji Watanabe, and Takaaki Hori, “Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling,” in Proc. of SLT, 2018.

[8] Sibo Tong, Philip N Garner, and Hervé Bourlard, “Multilingual training and cross-lingual adaptation on CTC-based acoustic model,” arXiv preprint arXiv:1711.10025, 2017.

[9] Sibo Tong, Philip N Garner, and Hervé Bourlard, “Cross-lingual adaptation of a CTC-based multilingual acoustic model,” Speech Communication, vol. 104, pp. 39–46, 2018.

[10] Martin Karafiát, Murali Karthick Baskar, Karel Veselý, František Grézl, Lukáš Burget, et al., “Analysis of multilingual blstm acoustic model on low and high resource languages,” in Proc. of ICASSP, 2018, pp. 5789–5793.

[11] Karel Veselý, Martin Karafiát, František Grézl, Milos Janda, and Ekaterina Egorova, “The language-independent bottleneck features,” in Proc. of SLT, 2012, pp. 336–341.

[12] František Grézl and Martin Karafiát, “Adapting multilingual neural network hierarchy to a new language,” in Spoken Language Technologies for Under-Resourced Languages, 2014.

[13] Tatsuya Kawahara, “Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition,” in Proc. of SLT, 2018, to appear.

[14] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. of ICASSP, 2016, pp. 4960–4964.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proc. of ICML, 2006, pp. 369–376.

[16] Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, and Tatsuya Kawahara, “Leveraging sequence-to-sequence speech synthesis for neural machine translation through multi-task learning,” in Proc. of ICASSP, 2017, pp. 4835–4839.

[17] Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath, “Speech recognition and keyword spotting for low-resource languages: BEL project research at CUED,” in Spoken Language Technologies for Under-Resourced Languages, 2014.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in Proc. of ICASSP, 2015, pp. 5206–5210.

[19] Shinji Watanabe, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu, “A comparison of techniques for language model integration in encoder-decoder speech recognition,” arXiv preprint arXiv:1807.10857, 2018.

[20] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barraut, Hieu-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “On using monolingual corpora in neural machine translation,” arXiv preprint arXiv:1503.03535, 2015.

[21] Anuroop Sridram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates, “Cold fusion: Training seq2seq models together with language models,” in Proc. of Interspeech, 2018, pp. 387–391.

[22] Prajit Ramachandran, Peter J Liu, and Quoc V Le, “Unsupervised pre-training for sequence to sequence learning,” in Proc. of EMNLP, 2017.

[23] Tobias Domhan and Felix Hieber, “Using target-side monolingual data for neural machine translation through multi-task learning,” in Proc. of EMNLP, 2017, pp. 1500–1505.

[24] Thomas Lehnert, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda, “Back-translation-style data augmentation for end-to-end ASR,” in Proc. of SLT, 2018, to appear.

[25] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barraut, Hieu-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “Cold fusion: Training seq2seq models together with language models,” in Proc. of Interspeech, 2018, pp. 387–391.

[26] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Chen, and Rohit Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in Proc. of ICASSP, 2017, pp. 5824–5828.

[27] Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu, “A comparison of techniques for language model integration in encoder-decoder speech recognition,” arXiv preprint arXiv:1807.10857, 2018.