Vol.2020-SLP-133 No.8
2020/10/22

Super Multi-lingual End-to-End Speech Recognition and Its Transfer learning to Low Resource Languages

Wenxin Hou[1] Yue Dong[1] Bairong Zhuang[1] Longfei Yang[1] Jiatong Shi[2] Takahiro Shinozaki[1]

概要：We report the evaluation results of large-scale end-to-end multilingual speech recognition that recognizes 42 languages. The system is a hybrid CTC/attention architecture, and we train it using around 5000 hours of data. The average character error rate for the 42 languages was 27.8. Furthermore, we transfer the pre-trained model to 14 low-resource languages. Results show that the pre-trained model achieves significantly better results than a non-pretrained baseline.

1. Introduction

End-to-end automatic speech recognition (ASR) has shown great potential for high extensibility and flexibility. Moreover, in recent years the end-to-end model has attained comparable performance to conventional Hidden Markov Model-based systems [1]. However, it remains a problem to improve recognition performance on low-resource languages due to the scarcity of data. Cho et al. reported the effectiveness of transfer learning by pre-training on 10 source languages and then fine-tuning on 4 low-resource languages [2]. In this work, we investigated into super multilingual pre-training by training a single end-to-end model on 42 languages. Then we transferred the model to 14 low-resource languages and discussed the differences among languages [5].

2. Multilingual Speech Recognition

2.1 Language-Independent Architecture

To perform multilingual ASR, we adopted a language-independent architecture [3]. The model vocabulary shares characters across all the target languages. To reduce the possibility that predicted tokens switch between languages, we prepend corresponding language IDs (e.g. [en], [ja]) to the output target so that the model learns to predict the language ID then output the recognized text, which could be regarded as an auxiliary language identification task.

2.2 Transfer Learning to Low-resource Languages

We transfer the pre-trained model to low-resource languages under two settings: monolingual transfer and multilingual transfer. The output layers of CTC and decoder are replaced and all the parameters are then fine-tuned on limited labeled data of target language(s). Under the condition of multilingual transfer, we employ the same language-independent architecture as introduced in Section 2.1.

3. Experiments

3.1 Setup

Speech data of 42 languages from 11 corpora are used for pre-training: AISHELL, Aurora4, Babel, Common Voice, CSJ, CHiME4, Fisher Callhome, Fisher Switchboard, Voxforge, WSJ, HKUST). The total number of utterances are over 6 million and total duration is around 5,000 hours. We used 14 target languages from Common Voice as transfer targets as shown in Table 1. It is worth noting that 12 languages out of them are shorter than 10 hours, among which Kinyarwanda has only 0.25-hour data, while the longest Esperanto data are 35 hours. In our experiments, we randomly split 80% for training and 20% for testing.

80-dimensional filter-bank and 3-dimensional pitch features are extracted as model input. We used the

[1] 東京工業大学
Tokyo Institute of Technology, Tokyo, Japan
www.ts.ip.titech.ac.jp
[2] Johns Hopkins University

表 1 Low-resource languages

Language	Duration
Arabic	7
Breton	5
Hakka Chins	2
Chuvash	0.96
Dhivehi	6
Esperanto	35
Estonian	10
Indonesian	3
Interlingua	1
Kinyarwanda	0.25
Kyrgyz	11
Latvian	4
Sakha	3
Slovenian	3

Transformer-based hybrid CTC/Attention architecture. The detailed Transformer architecture follows the *big model* as explained in [4]. As the baseline for comparison, we trained a randomly-initialized model directly on target low-resource languages. All the experiments were performed on TSUBAME 3.0 supercomputer [*1].

3.2 Results

Table 2 presents the evaluation results of monolingual transfer. Due to the limitation of the data amount, the baseline model obtained high character error rates (CER). For Chuvash and Kinyarwanda which has extremely small data, CER went beyond 50%. On the other hand, the baseline achieves a relatively lower CER of 5.2% on Esperanto. Moreover, it can be observed that most of the languages witness a significant CER reduction by applying pre-training, demonstrating its effectiveness. However, probably as a consequence of extremely low data amount, CER was conversely increased with pre-training on Chuvash and Kinyarwanda. The average CER of 14 languages without pre-learning was 23%, while with pre-learning it was 9.6%, a large relative error reduction of 58.3% was obtained.

Table 3 presents CER results of multilingual transfer. The baseline model without pre-training was randomly initialized and trained on 14 low-resource language data simultaneously. Due to the use of multilingual data, significantly lower CER was obtained on all languages compared to those of the monolingual baseline. Compared with the non-pretrained model, CER was reduced with pre-learning in all 14 languages, including Chuvash and Kinyarwanda. It could be considered that multilingual fine-tuning stabilizes the transfer process for extremely low-resource data. On the other hand, by comparing the results of the monolingual and multilingual transfer, we notice that the latter setting slightly increases CER on some languages (e.g., Arabic). This could be considered as a side effect of learning languages other than the recognition target language at the same time.

[*1] https://www.t3.gsic.titech.ac.jp

表 2 Monolingual transfer

Language	w/o pre-training	w/ pre-training
Arabic	19.7	**15.7** (20.3%↓)
Breton	21.1	**15.8** (25.1%↓)
Hakha Chin	14.7	**10.0** (32.0%↓)
Chuvash	19.2	**14.4** (25.0%↓)
Dhivehi	13.7	**10.8** (21.2%↓)
Esperanto	3.8	**2.7** (28.9%↓)
Estonian	14.5	**10.1** (30.3%↓)
Indonesian	14.5	**10.1** (30.3%↓)
Interlingua	14.4	**10.5** (27.1%↓)
Kyrgyz	11.1	**8.0** (28.2%↓)
Latvian	13.7	**10.2** (25.5%↓)
Kinyarwanda	45.3	**31.7** (30.0%↓)
Sakha	15.3	**11.9** (22.2%↓)
Slovenian	10.7	**8.5** (20.6%↓)
Weighted Avg.	11.1	**8.2** (26.1%↓)

表 3 Multilingual transfer

Language	w/o pre-training	w/ pre-training
Arabic	19.7	**15.7** (20.3%↓)
Breton	21.1	**15.8** (25.1%↓)
Hakha Chin	14.7	**10.0** (32.0%↓)
Chuvash	19.2	**14.4** (25.0%↓)
Dhivehi	13.7	**10.8** (21.2%↓)
Esperanto	3.8	**2.7** (28.9%↓)
Estonian	14.5	**10.1** (30.3%↓)
Indonesian	14.5	**10.1** (30.3%↓)
Interlingua	14.4	**10.5** (27.1%↓)
Kyrgyz	11.1	**8.0** (28.2%↓)
Latvian	13.7	**10.2** (25.5%↓)
Kinyarwanda	45.3	**31.7** (30.0%↓)
Sakha	15.3	**11.9** (22.2%↓)
Slovenian	10.7	**8.5** (20.6%↓)
Weighted Avg.	11.1	**8.2** (26.1%↓)

4. Conclusion

We investigated into low-resource speech recognition by super multilingual pre-learning. It was confirmed that the recognition performance in low-resource languages was greatly improved by pre-training. A future task is to further improve the performance of multilingual transfer learning while suppressing side effects.

参考文献

[1] Chiu, Chung-Cheng, et al.: *State-of-the-art speech recognition with sequence-to-sequence models*, ICASSP (2018).
[2] Cho, Jaejin, et al.: *Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling*, IEEE SLT (2018).
[3] Watanabe, Shinji, et al.: *Language independent end-to-end architecture for joint language identification and speech recognition*, ASRU (2017).
[4] Dong, Linhao, et al.: *Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition*, ICASSP (2018).
[5] Wenxin, Hou, et al.: *Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning*, INTERSPEECH (2020).