Analyzing ASR Pretraining for Low-Resource Speech-to-Text Translation

Mihaela C. Stoian, Sameer Bansal, Sharon Goldwater

School of Informatics, University of Edinburgh, UK
{c.mihaela.stoian, sameer.bansal}@ed.ac.uk, sgwater@inf.ed.ac.uk

ABSTRACT

Previous work has shown that for low-resource source languages, automatic speech-to-text translation (AST) can be improved by pretraining an end-to-end model on automatic speech recognition (ASR) data from a high-resource language. However, it is not clear what factors—e.g., language relatedness or size of the pretraining data—yield the biggest improvements, or whether pretraining can be effectively combined with other methods such as data augmentation. Here, we experiment with pretraining on datasets of varying sizes, including languages related and unrelated to the AST source language. We find that the best predictor of final AST performance is the word error rate of the pretrained ASR model, and that differences in ASR/AST performance correlate with how phonetic information is encoded in the later RNN layers of our model. We also show that pretraining and data augmentation yield complementary benefits for AST.

Index Terms— speech-to-text translation, transfer learning, pretraining, speech recognition, data augmentation.

1. INTRODUCTION

Low-resource automatic speech-to-text translation (AST) has recently gained traction as a way to bring NLP tools to under-represented languages. An end-to-end approach [1,7] is particularly appealing for source languages with no written form, or for endangered languages where translations into a high-resource language may be easier to collect than transcriptions [8]. However, building high-quality end-to-end AST with little parallel data is challenging, and has led researchers to explore how other sources of data could be used to help. A number of methods have been investigated. Several of these use transcribed source language audio and/or translated source language text in a multitask learning scenario [4,9] or to pre-train parts of the model before fine-tuning on the end-to-end AST task [1]. Others assume, as we do here, that no additional source language resources are available, in which case transfer learning using data from language(s) other than the source language is a good option. In particular, several researchers have shown that low-resource AST can be improved by pretraining on an ASR task in some other language, then transferring the encoder parameters to initialize the AST model. For example, Bansal et al. [5] showed that pretraining on either English or French ASR improved their Spanish-English AST system (trained on 20 hours of parallel data) and Tian [10] got improvements on an 8-hour Swahili-English AST dataset using English ASR pretraining.

Overall these results show that pretraining helps, but leave open the question of what factors affect the degree of improvement. For example, does language relatedness play a role, or simply the amount of pretraining data? Bansal et al. showed bigger AST gains as the amount of English pretraining data increased from 20 to 300 hours, and also found a slightly larger improvement when pretraining on 20 hours of English versus 20 hours of French, but they pointed out that the Spanish data contains many English code-switched words, which could explain the latter result. In related work on multilingual pretraining for low-resource ASR, Adams et al. [11] showed that pre-training on more languages helps, but it is not clear whether the improvement is due to including more languages, or just more data.

To begin to tease apart these issues, we focus here on monolingual pretraining for low-resource AST, and investigate two questions. First, can we predict what sort of pretraining data is best for a particular AST task? Does it matter if the pretraining language is related to the AST source language (defined here as part of the same language family, since phonetic similarity is difficult to measure), or is the amount of pretraining data (or some other factor) more important? Second, can pretraining be effectively combined with other methods, such as data augmentation, in order to further improve AST results?

To answer these questions, we use the same AST architecture and Spanish-English parallel data as Bansal et al. [5], but pretrain the encoder using a number of different ASR datasets: the 150-hour AISHELL corpus of Chinese as well as seven GlobalPhone languages, each with about 20 hours of data. We find that pretraining on a larger amount of data from an unrelated language is much better than pretraining on a smaller amount of data from a related language. Moreover, even when controlling for the amount of data, the WER of the ASR model from pretraining seems to be a better predictor of final AST performance than does language relatedness. Indeed, we show that there is a very strong correlation between the WER of the pretrained model and BLEU score of the final AST model—i.e., the best pretraining strategy may simply be to use datasets and methods that will yield the lowest ASR WER during pretraining. However, we also found that AST results can be improved further by augmenting the AST data using standard speed perturbation techniques [12]. Our best results using non-English pretraining data improve the test set BLEU scores of an AST system trained on 20 hours of parallel data from 10.2 to 14.3, increasing to 15.8 with data augmentation.

Finally, we analyze the representations learned by the models and show that better performance seems to correlate with the extent to which phonetic information is encoded in a linearly separable way in the later RNN layers.

2. METHODOLOGY

For both ASR and AST tasks we use the same end-to-end system architecture shown in Figure 1: the encoder-decoder model from [5], which itself is adapted from [2,4] and [3]. Details of the architecture and training parameters are described in Section 3.4.
3. EXPERIMENTAL SETUP

3.1. Parallel data

For the AST models, we use Spanish-English parallel data from Fisher corpus [15], containing 160 hours of Spanish telephone speech translated into English text. To simulate low-resource settings, we randomly downsample the original corpus to 20 hours of training data. Each of the dev and test sets comprise 4.5 hours of speech.

3.2. Pretraining data

Since we focus on investigating factors that might affect the AST improvements over the baseline when pretraining, we have chosen ASR datasets for pretraining that contrast in the number of hours (WER).

To explore the effects of using a large amount of pretraining data from an unrelated language, we used the AISHELL-1 corpus of Mandarin Chinese [17], which contains 150 hours of read speech. Transcriptions with annotated word boundaries are available in both Hanzi (Chinese characters) and Romanized versions, and we built models with each. To compare to the GlobalPhone data, we also created a 20-hour subset of the Romanized AISHELL (zh-ai-small) by randomly selecting utterances from a subset of the speakers (81, roughly the number present in most of the GlobalPhone datasets).

Finally, to reproduce one of the experiments from [5], we pre-trained one model using 300 hours of Switchboard English [18]. This data is the most similar to the AST speech data in terms of style and channel (both are conversational telephone speech). However, as noted by [5], the Fisher Spanish speech contains many words that are actually in English (code-switching), so pretraining on English may provide an unfair advantage relative to other languages.

3.3. Preprocessing

We compute 13-dim MFCCs and cepstral mean and variance normalization along speakers using Kaldi’s [13] 3-way speed perturbation, adding versions of the AST data where the audio is sped down and up by a factor of 0.9 and 1.1, respectively. To evaluate ASR performance we compute the word error rate (WER). To evaluate AST performance we calculate the 4-gram BLEU score [14] on four reference translations.

To account for unseen words in the test data, we model the ASR data with unique word tokens, and from the final AST system (right). AST results in all rows except the first are from pretraining using the dataset listed in that row, followed by fine-tuning using ast-20h. Numbers in brackets are the improvement over the baseline.

To explore the effects of using a large amount of pretraining data from an unrelated language, we used the AISHELL-1 corpus of Mandarin Chinese [17], which contains 150 hours of read speech. Transcriptions with annotated word boundaries are available in both Hanzi (Chinese characters) and Romanized versions, and we built models with each. To compare to the GlobalPhone data, we also created a 20-hour subset of the Romanized AISHELL (zh-ai-small) by randomly selecting utterances from a subset of the speakers (81, roughly the number present in most of the GlobalPhone datasets).

Finally, to reproduce one of the experiments from [5], we pre-trained one model using 300 hours of Switchboard English [18]. This data is the most similar to the AST speech data in terms of style and channel (both are conversational telephone speech). However, as noted by [5], the Fisher Spanish speech contains many words that are actually in English (code-switching), so pretraining on English may provide an unfair advantage relative to other languages.

3.4. Model architecture and training

Following the architecture and training procedure described in [5], input speech features are fed into a stack of two CNN layers. In each CNN layer we stride the input with a factor of 2 along time, apply
ReLU activation [20] followed by batch normalization [21]. The CNN output is fed into a three-layer bi-directional long short-term memory network (LSTM) [22], with 512 hidden layer dimensions. For decoding, we use the predicted token 20% of the time and the training token 80% of the time [23] as input to a 128-dimensional embedding layer followed by a three-layer LSTM, with 256 hidden layer dimensions, and combine this with the output from the attention mechanism [24] to predict the word at the current time step.

We use code and hyperparameter settings from [5] the Adam optimizer [25] with an initial learning rate of 0.001 and decay it by a factor of 0.5 based on the dev set BLEU score. When training AST models, we regularize using dropout [26] with a ratio of 0.3 over the embedding and LSTM layers [27], weight decay with a rate of 0.0001; and, after the first 20 epochs, 30% of the time we replace the predicted output word by a random word from the target vocabulary. At test time we use beam decoding with a beam size of 5 and length normalization [28] with a weight of 0.6.

4. RESULTS AND DISCUSSION

4.1. Baseline and ASR results

Our baseline 20-hour AST system obtains a BLEU score of 10.3 (Table 1 first row), 0.5 BLEU point lower than that reported by [5]. This discrepancy might be due to differences in subsampling from the 160-hour AST dataset to create the 20-hour subset, or from Kaldi parameters when computing the MFCCs.

WERs for our pre-trained models (Table 1) vary from 22.5 for the large AISHELL dataset with Romanized transcript to 80.5 for Portuguese GlobalPhone. These are considerably worse than state-of-the-art ASR systems (e.g., Kaldi recipes can achieve WER of 7.5 on AISHELL and 26.5 on Portuguese GlobalPhone), but we did not optimize our architecture or hyperparameters for the ASR task since our main goal is to analyze the relationship between pretraining and AST performance (and in order to use pretraining, we must use a seq2seq model with the architecture as for AST).

4.2. Pretraining the AST task on ASR models

AST results for our pre-trained models are given in Table 1. Pretraining improves AST performance in every case, with improvements ranging from 0.2 (pt-gp) to 4.3 (zh-ai-large). These results make it clear that language relatedness does not play a strong role in predicting ASR improvements, since on the similar-sized GlobalPhone datasets, the two languages most related to Spanish (French and Portuguese) yield the highest and lowest improvements, respectively. Moreover, pretraining on the large Chinese dataset yields a bigger improvement than either of these—4.3 BLEU points. This is nearly as much as the 6 point improvement reported by [5] when pretraining on 100 hours of English data, which is especially surprising given not only that Chinese is very different from Spanish, but also that the Spanish data contains some English words.

This finding seems to suggest that data size is more important than language relatedness for predicting the effects of pretraining. However, there are big differences even amongst the languages with similar amounts of pretraining data. Analyzing our results further, we found a striking correlation between the WER of the initial ASR model and the BLEU score of the AST system pretrained using that model, as shown in Figure 2. Therefore, although pretraining data size clearly influences ASR performance, this appears to be mainly due to its effect on WER of the ASR model. We therefore hypothesize that WER is a better direct predictor of AST performance than either data size or language relatedness.

4.3. Multilingual pretraining

Although our main focus is monolingual pretraining, we also looked briefly at multilingual pretraining, inspired by recent work on multilingual ASR [29, 30] and evidence that multilingual pretraining followed by fine-tuning on a distinct target language can improve ASR on the target language [11, 31, 32]. These experiments did not directly compare pretraining using a similar amount of monolingual data, but such a comparison was done by [33, 34] in their work on learning feature representations for a target language with no transcribed data. They found a benefit for multilingual vs monolingual pretraining given the same amount of data.

Following up on this work, we tried pretraining using 124 hours of multilingual data (all GlobalPhone languages except Chinese), roughly the amount of data in our large Chinese models. We combined all the data together and trained an ASR model using a common target BPE with 6k merge operations, then transferred only the encoder to the AST model. However, we did not see a benefit to the multilingual training (Table 1 final row); in fact the resulting AST model was slightly worse than the zh-ai-large model (BLEU of 13.3 vs 14.6). Other configurations of multilingual training might still outperform their monolingual counterparts, but we leave this investigation as future work.

4.4. Augmenting the parallel data

Table 2 (top) shows how data augmentation affects the results of the baseline 20h AST system, as well as three of the best-performing pretrained models from Table 1. For these experiments only, we changed the learning rates of the augmented-data systems so that all models took about the same amount of time to train (see Figure 3). Despite a more aggressive learning schedule, the performance of the augmented-data systems surpasses that of the baseline and pretrained models, even those trained on the largest ASR sets (150-hr Chinese and 300-hr English).

For comparison to other work, Table 2 (bottom) gives results for AST models trained on the full 160 hours of parallel data, including models with both pretraining and data augmentation. For the latter, we used the original learning schedule, but had to stop training early due
5. ANALYZING THE MODELS’ REPRESENTATIONS

Finally, we hope to gain some understanding into why pretraining on ASR helps with AST, and specifically how the neural network representations change during pretraining and fine-tuning. We follow [35] and [10], who built diagnostic classifiers [36] to examine the representation of phonetic information in end-to-end ASR and AST systems, respectively. Unlike [10,35], who used non-linear classifiers, we use a linear classifier to predict phone labels from the internal representations of the trained ASR or AST model.

Using a linear classifier allows us to make more precise claims: if the classifier performs better using the representation from a particular layer, we can say that layer represents the phonetic information in a more linearly separable way. Using a nonlinear classifier raises questions about how to choose the complexity of the classifier itself, and therefore makes any results difficult to interpret.

We hypothesized that pretraining allows the models to abstract away from nonlinguistic acoustic differences, and to better represent phonetic information: crucially, both in the trained language and in other languages. To test this hypothesis, we used two phone-labelled datasets distinct from all our ASR and AST datasets: the English TIMIT corpus (a language different to all of our trained models, with hand-labeled phones) and the Spanish GlobalPhone corpus (the same language as our AST source language, with phonetic forced-alignments produced using Kaldi). We randomly sampled utterances from these and passed them through the trained encoders, giving us a total of about 600k encoded frames. We used 400k of these to train logistic regression models to predict the phone labels, and tested on the remaining 200k frames.

Separate logistic regression models were trained on the representations from each layer of the encoder. Since convolutional layers have a stride of 2, the number of frames decreases at each convolutional layer. To label the frames after a convolutional layer we eliminated every other label (and corresponding frame) from the original label sequence. For example, given label sequence $S_1 = \text{aaaanaam}$ at input layer, we get sequence $S_2 = \text{aan}$ at the first convolutional layer and sequence $S_3 = \text{aan}$ at the second convolutional layer and at the following recurrent layers.

Results for the two classification data sets (Figure 4) show very similar patterns. In both the ASR and the AST models, the pretraining data seems to make little difference to phonetic encoding at the early layers, and classification accuracy peaks at the second CNN layer. However, the RNN layers show a clear trend where phone classification accuracy drops off more slowly for models with better ASR/AST performance (i.e., zh $>$ fr $>$ pt). That is, the later RNN layers more transparently encode language-universal phonetic information.

Phone classification accuracy in the RNN layers drops for both English and Spanish after fine-tuning on the AST data. This is slightly surprising for Spanish, since the fine-tuning data (unlike the pretraining data) is actually Spanish speech. However, we hypothesize that for AST, higher layers of the encoder may be recruited more to encode semantic information needed for the translation task, and therefore lose some of the linear separability in the phonetic information. Nevertheless, we still see the same pattern where better end-to-end models have higher classification accuracy in the later layers.

6. CONCLUSIONS

This paper explored what factors help pretraining for low-resource AST. We performed careful comparisons to tease apart the effects of language relatedness and data size, ultimately finding that rather than either of these, the WER of the pre-trained ASR model is likely the best direct predictor of AST performance. Given equivalent amounts of data, we did not find multilingual pretraining to help more than monolingual pretraining, but we did find an added benefit from using speed perturbation to augment the AST data. Finally, analysis of the pretrained models suggests that those models with better WER are transparently encoding more language-universal phonetic information in the later RNN layers, and this appears to help with AST.

Table 2: BLEU scores on dev and test sets for models trained with and without data augmentation. We used either 20h of AST training data (top block) or 160h (bottom block), with various pretraining.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20h –</td>
<td>10.3</td>
<td>13.0 (+2.7)</td>
<td>10.2</td>
<td>13.3 (+3.1)</td>
</tr>
<tr>
<td>fr-gp</td>
<td>12.5</td>
<td>13.7 (+1.2)</td>
<td>12.6</td>
<td>14.3 (+1.7)</td>
</tr>
<tr>
<td>zh-ai-lg</td>
<td>14.6</td>
<td>15.5 (+0.9)</td>
<td>14.3</td>
<td>15.8 (+1.5)</td>
</tr>
<tr>
<td>en-300h</td>
<td>19.5</td>
<td>20.1 (+0.6)</td>
<td>20.1</td>
<td>20.2 (+0.1)</td>
</tr>
<tr>
<td>160h –</td>
<td>34.1</td>
<td>36.3 (+2.2)</td>
<td>34.6</td>
<td>37.3 (+2.7)</td>
</tr>
<tr>
<td>en-300h</td>
<td>36.3</td>
<td>37.9 (+1.6)</td>
<td>36.4</td>
<td>37.8 (+1.4)</td>
</tr>
</tbody>
</table>

Fig. 4: Phonetic classification accuracy at different layers of our ASR (left) and AST (right) models. Different color bars indicate representations extracted from models (pre)trained on different datasets (pt-gp, fr-gp, or zh-ai-large). Results from the baseline AST model (without pretraining) are shown in both panels for comparison. The bars with black edges are results on TIMIT (majority baseline: 12.9%); the taller bars are for Spanish GlobalPhone (majority baseline: 15.2%).

Fig. 3: The AST performance over time (without beam-search) of baseline, pretrained, and pretrained+augmented models.
7. ACKNOWLEDGEMENTS

The authors wish to thank Yusheng Tian for her work on her Master’s thesis at the University of Edinburgh which inspired the analysis of the change in neural network representations during pretraining and fine-tuning. Also, thanks to Dr. Yevgen Matusevych and to Ramon Sanabria for useful discussions, proof-reading and providing feedback on the paper. This work was supported in part by a James S. McDonnell Foundation Scholar Award (220020374).

8. REFERENCES


