Towards Language-Universal End-to-End Speech Recognition

Suyoun Kim\textsuperscript{1}, and Michael L. Seltzer\textsuperscript{2}

\textsuperscript{1}Carnegie Mellon University
\textsuperscript{2}Facebook (formerly Microsoft AI & Research)

Microsoft
Carnegie Mellon University

ICASSP April 18, 2018
Presenter: Suyoun Kim
Outline

• Motivation of Language-universal end-to-end speech recognition
• Proposed model: language-specific gated network
• Experimental evaluation
• Conclusions
Challenges of growing language coverage of ASR systems

• There are over 6,000 languages globally

1) Conventional ASR requires each model be trained independently
   • Effort to train, deploy, and maintain so many models in production increases

2) For second and third tier languages, additional challenges arise
   • Lack of sufficient training data
   • Lack of linguistic expertise, lexicons
Prior work: multi-lingual acoustic models

• Transfer learning approach:
  • Share language-independent lower layer(s)
  • Separate language-specific output layer(s)

✓ Pools data to train common parameters
✓ Improved performance with (very) little training data

✗ Requires pronunciation lexicon
✗ Improvement diminishes with increased data
Our model: A language-universal end-to-end ASR

• Key insights

1) End-to-end with CTC
2) Universal character set
3) Language-specific gating
Our model: A language-universal end-to-end ASR

• Key insights

1) End-to-end with CTC¹

• No pronunciation lexicon required

• Convert a sequence of features to a sequence of graphemes rather than senones

¹ Graves et al. 2016
Our model: A language-universal end-to-end ASR

• Key insights

1) End-to-end with CTC
   • No pronunciation lexicon required

2) Universal character set
   • Single system
   • Easy to maintain

3) Language-specific gating
2) Use a universal character set

- Share model parameters and even output layer among languages
  - **Single system** capable of recognizing any language it has been trained on

- Assume language identity is known in training and decoding

- Mask out the activation from unwanted characters

Union Labels

“Universal keyboard” shares common characters
Experiment setup

• Data
  • Cortana data in English (EN), Spanish (ES), and German (DE)
  • 150 hour training set, 10 hour dev set, 10 hour test set, per language

• Model:
  • Input: 80-dimensional log mel filterbank x 3
  • Output: characters (graphemes)\(^1\) - EN: 81d, DE: 93d, ES: 97d
  • 4 layer BLSTM (320 cells)

• Training and Decoding
  • CTC with SGD with fixed learning rate, early stopping, random initialization
  • Greedy decoding with no explicit language model

\(^1\) Zweig et al., advances in all-neural speech recognition, 2016
### Initial evaluation:

<table>
<thead>
<tr>
<th>Training Languages</th>
<th>Total Hrs</th>
<th>Model Arch</th>
<th>Test Lang</th>
<th>CER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>150</td>
<td>mtl</td>
<td></td>
<td>23.3</td>
</tr>
<tr>
<td>DE + EN</td>
<td>300</td>
<td>mtl</td>
<td></td>
<td>22.3</td>
</tr>
<tr>
<td>DE + EN</td>
<td>300</td>
<td>univ</td>
<td>DE</td>
<td>22.5</td>
</tr>
<tr>
<td>DE + EN + ES</td>
<td>450</td>
<td>univ</td>
<td>DE</td>
<td>22.8</td>
</tr>
<tr>
<td>DE</td>
<td>300</td>
<td></td>
<td></td>
<td>15.8</td>
</tr>
<tr>
<td>ES</td>
<td>150</td>
<td>mtl</td>
<td></td>
<td>13.7</td>
</tr>
<tr>
<td>ES + EN</td>
<td>300</td>
<td>mtl</td>
<td></td>
<td>13.1</td>
</tr>
<tr>
<td>ES + EN</td>
<td>300</td>
<td>univ</td>
<td>ES</td>
<td>12.9</td>
</tr>
<tr>
<td>ES + EN + DE</td>
<td>450</td>
<td>univ</td>
<td></td>
<td>13.1</td>
</tr>
<tr>
<td>ES</td>
<td>300</td>
<td></td>
<td></td>
<td>11.7</td>
</tr>
</tbody>
</table>

1. Small gain by adding different EN training source
Initial evaluation: multi-task vs. union architectures

<table>
<thead>
<tr>
<th>Training Languages</th>
<th>Total Hrs</th>
<th>Model Arch</th>
<th>Test Lang</th>
<th>CER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>150</td>
<td></td>
<td></td>
<td>23.3</td>
</tr>
<tr>
<td>DE + EN</td>
<td>300</td>
<td>mtl</td>
<td></td>
<td>22.3</td>
</tr>
<tr>
<td>DE + EN</td>
<td>300</td>
<td>univ</td>
<td>DE</td>
<td>22.5</td>
</tr>
<tr>
<td>DE + EN + ES</td>
<td>450</td>
<td>univ</td>
<td></td>
<td>22.8</td>
</tr>
<tr>
<td>DE</td>
<td>300</td>
<td></td>
<td></td>
<td>15.8</td>
</tr>
<tr>
<td>ES</td>
<td>150</td>
<td></td>
<td></td>
<td>13.7</td>
</tr>
<tr>
<td>ES + EN</td>
<td>300</td>
<td>mtl</td>
<td></td>
<td>13.1</td>
</tr>
<tr>
<td>ES + EN</td>
<td>300</td>
<td>univ</td>
<td>ES</td>
<td>12.9</td>
</tr>
<tr>
<td>ES + EN + DE</td>
<td>450</td>
<td>univ</td>
<td></td>
<td>13.1</td>
</tr>
<tr>
<td>ES</td>
<td>300</td>
<td></td>
<td></td>
<td>11.7</td>
</tr>
</tbody>
</table>

1. Small gain by adding different EN training source
2. Separate labels (mtl) and universal labels (univ) perform comparably
Initial evaluation:
No improvement increasing from 2 langs. to 3 langs.

<table>
<thead>
<tr>
<th>Training Languages</th>
<th>Total Hrs</th>
<th>Model Arch</th>
<th>Test Lang</th>
<th>CER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>150</td>
<td></td>
<td></td>
<td>23.3</td>
</tr>
<tr>
<td>DE + EN</td>
<td>300</td>
<td>mtl</td>
<td></td>
<td>22.3</td>
</tr>
<tr>
<td>DE + EN</td>
<td>300</td>
<td>univ</td>
<td>DE</td>
<td>22.5</td>
</tr>
<tr>
<td>DE + EN + ES</td>
<td>450</td>
<td>univ</td>
<td></td>
<td>22.8</td>
</tr>
<tr>
<td>DE</td>
<td>300</td>
<td></td>
<td></td>
<td>15.8</td>
</tr>
<tr>
<td>ES</td>
<td>150</td>
<td>mtl</td>
<td></td>
<td>13.7</td>
</tr>
<tr>
<td>ES + EN</td>
<td>300</td>
<td>mtl</td>
<td></td>
<td>13.1</td>
</tr>
<tr>
<td>ES + EN</td>
<td>300</td>
<td>univ</td>
<td>ES</td>
<td>12.9</td>
</tr>
<tr>
<td>ES + EN + DE</td>
<td>450</td>
<td>univ</td>
<td></td>
<td>13.1</td>
</tr>
<tr>
<td>ES</td>
<td>300</td>
<td></td>
<td></td>
<td>11.7</td>
</tr>
</tbody>
</table>

1. Small gain by adding different EN training source
2. Separate labels (mtl) and universal labels (univ) perform comparably
3. No improvement increasing from 2 languages to 3 languages
Our model: A language-universal end-to-end ASR

• Key insights

1) End-to-end with CTC
   • No pronunciation lexicon required

2) Universal character set
   • Single system
   • Easy to maintain

3) Language-specific gating
   • Further improvement with more data
3) language-specific gating

• Motivation: model needs to adequately capture language-specific information
  • Adding language ID indicator gives minimal improvement

=> Add language-specific gating mechanism to each layer
  • Modulate internal representations in a language-specific way
  • Fewer parameter than *cluster adaptive training (CAT)*\(^1\)\(^2\)

---

\(^1\) Li et al., *multi-dialect speech recognition with a single sequence-to-sequence model*, 2018
\(^2\) Tan et al., *cluster adaptive training for deep learning network based acoustic model*, 2016
3) **language-specific gating: implementation details**

1. Define one-hot language indicator vector $d_l$
   
   $$d_l = [0 \ 0 \ 1]$$

2. Compute gate for $i^{th}$ hidden layer
   
   $$g(h_i, l) = \sigma(Uh_i + Vd_l + b)$$

3. Compute language-gated activation
   
   $$\hat{h}_i = g(h_i, l) \odot h_i$$

4. Gated activations and $d_i$ input to next layer
   
   $$\tilde{h}_i = [\hat{h}_i \colon d_i]$$
EN evaluation:
10.7% rel. impr. in CER, 7.0% rel. impr. in WER

- **without Gate**, no benefit increasing from 2 languages to 3 languages
- **with Gate**, additional gain increasing from 2 languages to 3 languages
DE evaluation:
11.4% rel. impr. in CER, and 8.6% rel. impr. in WER

- **without Gate**, no benefit increasing from 2 languages to 3 languages
- **with Gate**, additional gain increasing from 2 languages to 3 languages
ES evaluation:
14.1% rel. impr. in CER, and 11.1% rel. impr. in WER

- **without Gate**, no benefit increasing from 2 languages to 3 languages
- **with Gate**, additional gain increasing from 2 languages to 3 languages
Different ways to add language information to the model

- Adding one-hot language ID input gives minimal improvement ( + 0.1M parameters)
- Proposed approach results in the largest improvement, ( + 0.4M parameters, much fewer than cluster adaptive training\textsuperscript{12})

\textsuperscript{1} Li et al., multi-dialect speech recognition with a single sequence-to-sequence model, 2018
\textsuperscript{2} Tan et al., cluster adaptive training for deep learning network based acoustic model, 2016
Language-universal model can be a good initial model for creating a language-specific model

<table>
<thead>
<tr>
<th>Initial Model</th>
<th>Fine Tune</th>
<th>DE CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>--</td>
<td>DE (150h)</td>
<td>23.3</td>
</tr>
<tr>
<td>EN (1000h)</td>
<td>DE (150h)</td>
<td>21.4</td>
</tr>
<tr>
<td>EN + DE (300h)</td>
<td>DE (150h)</td>
<td>21.1</td>
</tr>
<tr>
<td>EN + ES + DE + gate (450h)</td>
<td>--</td>
<td>20.6</td>
</tr>
<tr>
<td>EN + ES + DE + gate (450h)</td>
<td>DE (150h)</td>
<td>19.4</td>
</tr>
</tbody>
</table>

- Fine-tuning DE from our universal model gets further gain - (5.8%)
- Our universal model is better initial model than EN (1000hr), well-trained monolingual from a different language - (9.3%)
Conclusions

• Our Language-Universal End-to-End Speech Recognition model

  • Does not require lexicon information and easy to maintain in production
  • Shows 7.0% - 11.1% WER reduction over monolingual character-based model
  • Shows 9.1% - 12.4% WER reduction over conventional MTL approach
  • Can be used as a good initial model for the further adaptation
    • Improves performance over bootstrapping from a well-trained monolingual from a different language
  • Need to evaluate with explicit language model