Multilingual and Low-Resource Speech Recognition

Peter Bell

Automatic Speech Recognition – ASR Lecture 15
8 March 2021
Languages of the World

- Over 6,000 languages globally....
- In Europe alone
  - 24 official languages and 5 “semi-official” languages
  - Over 100 further regional/minority languages
  - If we rank the 50 most used languages in Europe, then there are over 50 million speakers of languages 26-50 (Finnish – Montenegrin)
- 3,000 of the world’s languages are endangered
- Google cloud speech API covers over 98 languages and more than 300 accents/dialects of those languages; Apple Siri covers over 21 languages; Google assistant has over 30
Under-resourced (or low-resourced) languages have some or all of the following characteristics:

- limited web presence
- lack of linguistic expertise
- lack of digital resources: acoustic and text corpora, pronunciation lexica, ...

Under-resourced languages thus provide a challenge for speech technology.

See Besaciera et al (2014) for more.
Speech recognition of under-resourced languages

- Training acoustic and language models with limited training data
- Transferring knowledge between languages
- Constructing pronunciation lexica
- Dealing with language specific characteristics (e.g. morphology)
Many languages are morphologically richer than English: this has a major effect of vocabulary construction and language modelling.

**Compounding** (eg German): decompose compound words into constituent parts, and carry out pronunciation and language modelling on the decomposed parts.

**Highly inflected languages** (eg Arabic, Slavic languages): specific components for modelling inflection (eg factored language models).

**Inflecting and compounding languages** (eg Finnish).

All approaches aim to reduce ASR errors by reducing the OOV rate through modelling at the morph level; also addresses data sparsity.
Vocabulary size for different languages

Figure 7. Vocabulary growth curves for different languages: For growing amounts of text (word tokens), the numbers of unique different word forms (word types), occurring in the text are plotted.

3.3 Word Models, Vocabulary Growth, and Spontaneous Speech

To improve the word models, one could attempt to increase the vocabulary (recognition lexicon) of these models. A high coverage of the vocabulary of the training set might also reduce the OOV rate of the recognition data (test set). However, this may be difficult to obtain.

Figure 7 shows the development of the size of the training set vocabulary for growing amounts of training data. The corpora used for Finnish, Estonian, and Turkish are the datasets used for training language models (mentioned in Section 3.1.2). For comparison, a curve for English is also shown; the English corpus consists of text from the New York Times magazine. While there are fewer than 200,000 different word forms in the 40-million word English corpus, the corresponding values for Finnish and Estonian corpora of the same size exceed 1.8 million and 1.5 million words, respectively. The rate of growth remains high as the entire Finnish LM training data of 150 million words (used in Fin4) contains more than 4 million unique word forms. This value is thus ten times the size of the (rather large) word lexicon currently used in the Finnish experiments.

Figure 8 illustrates the development of the OOV rate in the test sets for growing amounts of training data. That is, assuming that the entire vocabulary of the training set is used as the recognition lexicon, the words in the test set that do not occur in the training set are OOVs. The test sets are the same as used in the speech recognition experiments, and for English, a held-out subset of the New York Times corpus was used. Again, the proportions of OOVs are fairly high for Finnish and Estonian; at 25 million words, the OOV rates are 3.6% and 4.4%, respectively (compared with 1.7% for Turkish and only 0.74%)

Creutz et al (2007)
OOV Rate for different languages

For growing amounts of training data, development of the proportions of words in the test set that are not covered by the training set.

For English. If the entire 150-million word Finnish corpus were to be used (i.e., a lexicon containing more than 4 million words), the OOV rate for the test set would still be 1.5%.

Not surprisingly, the feasibility of the use of high-coverage standard word lexicons for Finnish and Estonian is low. In light of the plots in Figures 7 and 8, word lexicons might, however, be an option for Turkish. The slower vocabulary growth for Turkish is likely due to the much lower number of compound words in Turkish in comparison to Finnish and Estonian. Word lexicons are the state-of-the-art solution for English.

3.3.1 Egyptian Arabic.

The vocabulary growth and OOV curves for Arabic are not visible in Figures 7 and 8 because of the small amount of Arabic data available (164,000 words). However, Figures 9 and 10 provide a close-up of the first 164,000 words, including Arabic. The datasets shown in Figures 7 and 8 all consist of planned, written text, whereas the ECA corpus contains unplanned, transcribed spontaneous speech. Because of these differences, the type of text (planned or spontaneous) has been indicated explicitly in the new figures.

Additional sources have been provided for Arabic and English: planned Arabic text from the FBIS corpus of Modern Standard Arabic (a collection of transcribed radio newscasts from various radio stations in the Arabic-speaking world) as well as spontaneous transcribed English telephone conversations from the Fisher corpus.

The point here is to illustrate that a smaller, slower growing vocabulary is used in spontaneous speech than in planned speech.

3 Available at http://www.ldc.upenn.edu/.

Creutz et al (2007)
Segmenting into morphs

- Linguistic rule-based approaches – require a lot of work for an under-resourced language!
- Automatic approaches – use automatically segment and cluster words into their constituent morphs
  - “Morfessor is an unsupervised data-driven method for the segmentation of words into morpheme-like units.”
  - Aims to identify frequently occurring substrings of letters within either a word list (type-based) or a corpus of text (token-based)
  - Uses a probabilistic framework to balance between few, short morphs and many, longer morphs
- Morph-based language modelling uses morphs instead of words – may require longer context (since multiple morphs correspond to one word)
Speech recognition accuracies on Finnish (Fin1-Fin4), Estonian (Est), Turkish (Tur), and Egyptian Arabic (ECA), using morph- and word-based n-gram language models.

Creutz et al (2007)
Multilingual and cross-lingual acoustic models

How to share information from acoustic models in different languages?

- General principle – use neural network hidden layers to learn a multilingual representation
- Hidden layers are multilingual – shared between languages
- Output layer are often monolingual language specific
- **Multi-lingual phone sets** use a network with multilingual hidden representations directly in a hybrid DNN/HMM systems
- **Hat-swap/multi-task** train a network with an output layer for each language, but shared hidden layers
- **Multilingual bottleneck** use a bottleneck hidden layer (trained in a multilingual) way as features for either a GMM- or NN-based system
Hat Swap – architecture

Fig. 1. Multilingual training of deep neural networks.

Does not require retraining any previously trained models for other languages. Ideally, one would like the hidden layers to converge to an optimized set of feature extractors that can be reused across domains and languages. However, such a study is inherently empirical, and variations of the techniques reported here are currently under investigation.

4. EXPERIMENTS

We used the GlobalPhone corpus for our experiments. The corpus consists of recordings of speakers reading newspapers in their native language. There are 19 languages from a variety of geographical locations: Asia (Chinese, Japanese, Korean), Middle East (Arabic, Turkish), Africa (Hausa), Europe (French, German, Polish), and Americas (Costa Rican Spanish, Brazilian Portuguese). Recordings are made under relatively quiet conditions using close-talking microphones; however, acoustic conditions may vary within a language and between languages.

In this work, we use seven languages from three different language families: Germanic, Romance, and Slavic. The languages used are: Czech, French, German, Polish, Brazilian Portuguese, Russian, and Costa Rican Spanish. Each language has roughly 20 hours of speech for training and two hours each for development and evaluation sets, from a total of about 100 speakers. The detailed statistics for each of the languages is shown in Table 1.

4.1. Baseline systems

For each language, we built standard maximum-likelihood (ML) trained GMM-HMM systems, using 39-dimensional MFCC features (C0-C12, with delta and acceleration coefficients), using the Kaldi speech recognition toolkit. The number of context-dependent triphone states for each language is 3100 with a total of 50K Gaussians (an average of roughly 16 Gaussians per state). The development set word error rates (WER) for the different languages are presented in Table 2. The results reported here are better than those in our earlier work because we used better LMs obtained from the authors of [3, 27]. We must stress that the ML baseline results are presented here to serve as a point of reference, and not for direct comparison with the DNN results. The scripts needed to replicate the GMM-HMM results are publicly available as a part of the Kaldi toolkit.

4.2. DNN configuration and results

For training DNNs, our tools utilize the Theano library, which supports transparent computation using both CPUs and GPUs. We train the networks on the same 39-dimensional MFCCs as the GMM-HMM baseline. The features are globally normalised to zero mean and unit variance, and 9 frames (4 on each side of the current frame) are used as the input to the networks. All the networks used here are 7 layers deep, with 2000 neurons per hidden layer. The initial weights for the softmax layer were chosen uniformly at random: $w \sim U[-r, r]$, where $r = \frac{4}{\sqrt{n_l+1}}$ and $n_l$ is the number of units in layer $l$. Fine-tuning is done using stochastic gradient descent on 256-frame mini-batches and an exponentially decaying schedule, learning at a fixed rate (0.08) until improvement in accuracy on cross-validation set between two successive epochs falls below 0.5%. The learning rate is then halved at each epoch until the overall accuracy fails to increase by 0.5% or more, at which point the algorithm terminates. While learning, the gradients were smoothed with 2Available from: http://kaldi.sf.net.

Table 1. Statistics of the subset of GlobalPhone languages used in this work: the amounts of speech data for training, development, and evaluation sets are in hours.

<table>
<thead>
<tr>
<th>Language</th>
<th>#Phones</th>
<th>#Spkrs</th>
<th>Train</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech (CZ)</td>
<td>41</td>
<td>102</td>
<td>26.8</td>
<td>2.4</td>
<td>2.7</td>
</tr>
<tr>
<td>French (FR)</td>
<td>38</td>
<td>100</td>
<td>22.8</td>
<td>2.1</td>
<td>2.0</td>
</tr>
<tr>
<td>German (DE)</td>
<td>41</td>
<td>77</td>
<td>14.9</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Polish (PL)</td>
<td>36</td>
<td>99</td>
<td>19.4</td>
<td>2.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Portuguese (PT)</td>
<td>45</td>
<td>101</td>
<td>22.8</td>
<td>1.6</td>
<td>1.8</td>
</tr>
<tr>
<td>Russian (RU)</td>
<td>48</td>
<td>115</td>
<td>19.8</td>
<td>2.5</td>
<td>2.4</td>
</tr>
<tr>
<td>Spanish (SP)</td>
<td>40</td>
<td>100</td>
<td>17.6</td>
<td>2.0</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Ghoshal et al, 2013
Recognition of GlobalPhone Polish

Ghoshal et al, 2013
Multi-lingual networks ("block softmax")

- Train one network for all languages:
  - separate output layer for each language
  - shared hidden layers
- Each training input is propagated forward to the output layer of the corresponding language – only that output layer is used to compute the error used to train the network for that input
- Since the hidden layers are shared, they must learn features relevant to all the output layers (languages)
- Can view this as a parallel version of hat swap
Multi-lingual networks – architecture

Huang et al, 2013

NB: A senone is a context-dependent tied state
Bottleneck features
Cross-lingual bottleneck features
Multi-lingual bottleneck network

Source language inputs

Source language outputs

Target language inputs

Target language outputs
## Multilingual bottleneck features – experiments

### GMM-based acoustic models

<table>
<thead>
<tr>
<th>Test language</th>
<th>WER [%]</th>
<th>MFCC</th>
<th>MFCC+BN Bottleneck trained on</th>
<th>MFCC+BN BN trained on</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>GER</td>
<td>ENU</td>
</tr>
<tr>
<td>GER</td>
<td>29.97</td>
<td>27.50 (8.2)</td>
<td>29.63 (1.1)</td>
<td>30.38 (-1.4)</td>
</tr>
<tr>
<td>ENU</td>
<td>21.69</td>
<td>21.31 (1.8)</td>
<td>18.85 (13.1)</td>
<td>22.63 (-4.3)</td>
</tr>
<tr>
<td>FRA</td>
<td>37.78</td>
<td>37.76 (0.1)</td>
<td>38.72 (-2.5)</td>
<td>33.95 (10.1)</td>
</tr>
</tbody>
</table>

(Mismatched acoustic environment)

Tüske et al, 2013
Use of BN features in HMM/DNN systems
Semi-supervised training

- Assume we only have a small amount of data is transcribed, but much more untranscribed data → train a seed model and use it to transcribe more data
- But don’t want to train further on incorrect captions
- Traditional solution: apply data filtering based on confidence scores
- This can select out the harder data that is most useful for refining the system
- Solution (Manohar, 2018): use a lattice to incorporate uncertainty about the transcription, train with LF-MMI criterion
Example: Tagalog
Self-supervised approaches

Conneau et al, 2020
Graphemes and phonemes

- Can represent pronunciations as a sequence of graphemes (letters) rather than a sequence of phones
- Advantages of grapheme-based pronunciations
  - No need to construct/generate phone-based pronunciations
  - Can use unicode attributes to assist in decision tree construction
- Disadvantages: not always direct link between graphemes and sounds (e.g. in English)
### Grapheme-based ASR results for 6 low-resource languages

<table>
<thead>
<tr>
<th>Language ID</th>
<th>System</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>tg</td>
</tr>
<tr>
<td>Kurmanji Kurdish 205</td>
<td>Phonetic Graphemic</td>
<td>67.6 67</td>
</tr>
<tr>
<td>Tok Pisin 207</td>
<td>Phonetic Graphemic</td>
<td>41.8 42.1</td>
</tr>
<tr>
<td>Cebuano 301</td>
<td>Phonetic Graphemic</td>
<td>55.5 55.5</td>
</tr>
<tr>
<td>Kazakh 302</td>
<td>Phonetic Graphemic</td>
<td>54.9 54.0</td>
</tr>
<tr>
<td>Telugu 303</td>
<td>Phonetic Graphemic</td>
<td>70.6 70.9</td>
</tr>
<tr>
<td>Lithuanian 304</td>
<td>Phonetic Graphemic</td>
<td>51.5 50.9</td>
</tr>
</tbody>
</table>

**IARPA Babel, 40h acoustic training data per language, monolingual training; cnc is confusion network combination, combining the grapheme- and phone-based systems**

Speech recognition systems for low-resource languages

- Morph-based language modeling
- Transferring data between acoustic models based on multilingual hidden representations
- Grapheme-based pronunciation lexica

In the future:
- “Zero-resource” ASR (no transcribed data at all)
- Languages without written forms
- Much active research in this area (including at Edinburgh)
Speech recognition systems for low-resource languages

- Morph-based language modeling
- Transferring data between acoustic models based on multilingual hidden representations
- Grapheme-based pronunciation lexica

In the future:
- “Zero-resource” ASR (no transcribed data at all)
- Languages without written forms
- Much active research in this area (including at Edinburgh)


