Pronunciation Lexicon Development for Under-Resourced Languages Using Automatically Derived Subword Units: A Case Study on Scottish Gaelic

Marzieh Razavi, Ramya Rasipuram and Mathew Magimai. Doss

28 November 2015
1 Motivation

2 Background

3 HMM-Based Formulation

4 Experimental Studies

5 Summary
Motivation
Standard speech technology systems model words as a sequence of subword units.

Using subword units necessitates availability of two resources:

1. The subword unit set.
   \[ A = \{a^1, \cdots, a^d, \cdots, a^D\} \]

2. The Lexicon mapping each word to a sequence of subword units.
Lexical Resources

- The most commonly used subword units: **Phones**
  - Linguistically motivated units: /f/, /p/, /b/, ···
  - Spectral envelope depicts the characteristics of phones.

- The phonetic lexicon provides the phonetic representation of words.

  phone → /f/ /ow/ /n/
  map → /m/ /ae/ /p/
  ···

- Typically developed manually.
- Augmented using grapheme-to-phoneme (G2P) conversion approaches.
- Require linguistic knowledge & human expertise.
Motivation

Lexicon Development for Under-Resourced languages

- Majority languages have well-developed lexicons.
- Under-resourced languages may lack proper lexical resources.
  - Examples: Uspanteko, Haitian Creole, . . .
  - Linguistic knowledge and human expertise may be very limited.
    ⇒ Conventional approaches cannot be exploited.

A possible solution

Automatically derive “phone-like” subword units and generate associated pronunciations using transcribed speech data.
Background
G2P Conversion

- Conventional Data-Driven G2P approaches:
  - Assumes the availability of a *seed lexicon* obtained using linguistic knowledge and expertise.
  - Apply machine learning techniques to learn the G2P relationship.

```
Word: phone

R='h'? 'p'
Y N

L='o'? R=consonant?
L='a'?
N Y
Y N
/p/ /f/ /p/ /p/ / € /
Y N

L='o'? R=consonant?
L='a'?
N Y
Y N
/p/ /f/ /p/ /p/ / € /
Y N
```

- Acoustic data-driven G2P conversion approach:
  - Assumes some speech data in addition to the seed lexicon is available.
  - G2P relationship is learned through acoustics.

R=Right-hand grapheme
L=Left-hand grapheme
Acoustic G2P Conversion Approach (I):
Learning the G2P Relationship Using Acoustics

**Step 1**
- **Acoustic observation sequence**
- **Categorical state distribution**
  \[ y_i^d = P(a^d|l_i) \]

**KL–HMM**
- \( l_i \): grapheme

**Step 2**
- **Phone posterior features**
  \[ z_t^d = P(a^d|x_t) \]
- \( a^d \): phone

**Acoustic observation sequence**
\[ (x_{t-4}, \ldots, x_t, \ldots, x_{t+4}) \]
\[ x_1, \ldots, x_t, \ldots, x_T \]
Acoustic G2P Conversion Approach (II): Pronunciation Inference Given the Learned G2P Relation

Word tokenizer

MAP
↓
\{M\} \{A\} \{P\}

\{M+A\} \{M-A+P\} \{A-P\}

Trained grapheme-based KL-HMM system

Phone posterior probability sequence

\begin{align*}
\text{Inferred pronunciation} &= [m \ ae \ p] \\
\text{Phone HMM decoder} &= \text{Ergodic HMM}
\end{align*}
HMM-Based Formulation
Formulating the Problem

- **Categorical state distribution**
  - $y_1^1, \ldots, y_D^1$
  - $y_1^2, \ldots, y_D^2$
  - $y_1^3, \ldots, y_D^3$

- **KL–HMM**
  - $l^1$: grapheme
  - $y_1^1, \ldots, y_D^1$
  - $y_1^2, \ldots, y_D^2$
  - $y_1^3, \ldots, y_D^3$

- **ASWU posterior features**
  - $z_1^1, \ldots, z_T^1$
  - $z_1^2, \ldots, z_T^2$
  - $z_1^3, \ldots, z_T^3$

- **Acoustic observation sequence**
  - $x_{t-4}, \ldots, x_t, \ldots, x_{t+4}$
  - $x_1, \ldots, x_t, \ldots, x_T$

- $\{a^d\}_{d=1}^D$ are automatic subword units (ASWUs) and unknown.
- Once $\{a^d\}_{d=1}^D$ discovered, apply the acoustic G2P conversion approach.
Standard HMM-based ASR

- Context-dependent (CD) subword units are modeled with HMMs with mixture of Gaussian state-output distributions.
- e.g. iphone → [i] [p] [h] [o] [n] [e] → [i+p] [i-p+h] [p-h+o] [h-o+n] [o-n+e] [n-e]

- Data sparsity issue
- Clustering and tying (parameter sharing) using decision trees:
  - maps each $G_{CD}$ to a $CS_G$

$G_{CD}$: context-dependent (CD) graphemes

$G_{CI}$: context independent (CI) graphemes

$CS_G$: clustered CD grapheme states (e.g. [ST_p_21])

$X$: acoustic cepstral features
Derivation of Automatic Phone-Like Subword Units

**Hypothesis:** The clustered CD grapheme states $CS_G$ can be treated as phone-like automatic subword units $a^d$.

- Cepstral feature $x$ carries phone-like information.
- $G_{CD}$ tends to relate to a phone in a regular manner.
  - Example:
    - $G_{CI} [p] \rightarrow /p/ , /f/$
    - $G_{CD} [p+h] \rightarrow /f/$
- $CS_G$ is found by maximizing the likelihood of the data.
- $CS_G$ relates to both $x$ and $G_{CD}$.

$\Rightarrow$ $CS_G$ should be phone-like.
**Proposed Approach**

Block Diagram of ASWU Derivation and Pronunciation Generation

(Phase I) Automatic subword unit derivation

- **Grapheme transcriptions**
- **Acoustic data**
  - DT-based Clustering
    - Clustered CD graphemes (ASWUs)

(Phase II) Learning the probabilistic G2ASWU relationship

- **Grapheme transcriptions**
- **ASWU posterior probabilities**
  - Grapheme-based KL-HMM
  - ANN

(Phase III) Pronunciation inference given the learned G2ASWU relationship

- **Text tokenizer**
  - CD grapheme sequence
  - Trained grapheme-based KL-HMM
  - ASWU posterior probability sequence
  - Ergodic HMM

Input word: AT

{A}{T} → {A,T} → {A+T}{A-T}

P([ST_A_21]) P([ST_A_21]) P([ST_A_21]) P([ST_A_21])

P([ST_Z_21]) P([ST_Z_21]) P([ST_Z_21]) P([ST_Z_21])

ASWU sequence: [ST_A_21] [ST_T_21] [ST_Z_21]
Experimental Studies
Scottish Gaelic

- Low-resourced and minority language; Endangered with only 60,000 speakers.
- Alphabet has 18 letters, consisting of 5 vowels and 13 consonants.
- There are 12 basic consonant types in Scottish Gaelic:
  - fortis or lenis: a grapheme [h] next to the consonant.
  - broad or slender: surrounded by ([a], [o], [u]) or ([e], [i])
- Number of graphemes is greater than number of phonemes in word.
  - an-diugh → /ə/ /n/ /d̪j/ /u/
  - aghaidh → /γ/ /i/
Under-Resourced Language Study: Scottish Gaelic

- Corpus was collected by the University of Edinburgh.
- Recordings from broadcast news and discussion program.
- Use transcribed speech data for subword unit derivation and pronunciation generation.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Lexicon size (in words)</th>
<th># of grapheme subword units</th>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scottish Gaelic</td>
<td>5082</td>
<td>31</td>
<td>180 (min) (22 speakers)</td>
<td>60 (min) (12 speakers)</td>
</tr>
</tbody>
</table>
Relating ASWUs to Phonetic Units

- No phonetic lexicon available.
- Exploit auxiliary linguistic resources from other languages.
Analysis of ASWU-based Pronunciations

- Map each ASWU to most probable multilingual phone.
- Provide the ‘perceived’ pronunciations for each word through informal hearing.

<table>
<thead>
<tr>
<th>Word</th>
<th>Lex-ASWU-82</th>
<th>mapped pron.</th>
<th>perceived pron.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHÀL</td>
<td>[ST_B_22] [ST_À_21] [S_L_23]</td>
<td>/v/ /a/ /l/</td>
<td>/v/ /a/ /l/</td>
</tr>
<tr>
<td>THOG</td>
<td>[ST_T_21] [ST_O_23] [ST_G_23]</td>
<td>/h/ /o/ /k/</td>
<td>/h/ /O/ /g/</td>
</tr>
<tr>
<td>PHÒS</td>
<td>[ST_F_21] [ST_Ò_21] [ST_S_23]</td>
<td>/f/ /o/ /s/</td>
<td>/f/ /o/ /s/</td>
</tr>
</tbody>
</table>

- The ASWU-based pronunciations to a certain extent capture the linguistic rules related to pronunciations.
ASR Study

<table>
<thead>
<tr>
<th>System</th>
<th># of units</th>
<th># of tied states</th>
<th>WA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM-GMM-GRAPH</td>
<td>32</td>
<td>1158</td>
<td>64.6</td>
</tr>
<tr>
<td>HMM-GMM-ASWU</td>
<td>82</td>
<td>1161</td>
<td>66.4</td>
</tr>
</tbody>
</table>

- **HMM-GMM-GRAPH**: Grapheme-based ASR
- **HMM-GMM-ASWU**: Proposed approach
- The ASWU-based lexicon yields a significantly better ASR system than the grapheme-based lexicon.
Summary

- Proposed an HMM formalism to derive phone-like subword units and generate associated pronunciations.
- The formalism is scalable to under-resourced languages.
- Investigated the potential of ASWUs for developing linguistically meaningful lexicons.
- Interpreted ASWUs in terms of linguistic units by exploiting auxiliary languages resources and prior linguistic knowledge.
Thank You