Methods for Efficient Semi-Automatic Pronunciation Dictionary Bootstrapping

Tim Schlippe, 18 August 2014
Outline

1. Introduction

2. Our semi-automatic pronunciation generation strategy

3. Experiments
   1. Analysis of intervals for G2P (grapheme-to-phoneme) model updates
   2. Use of single G2P converters
   3. Grapheme-to-phoneme converter combination
   4. Methods for initial grapheme-to-phoneme converter training data

4. Conclusion and future work
Motivation

- Pronunciation dictionaries needed for text-to-speech and automatic speech recognition (ASR)
- Manual production of pronunciations slow and costly
  - e.g. 19.2–30s / word for Afrikaans (Davel and Barnard, 2004)

→ Semi-automatic approaches reduce human effort but
  - are still time-consuming.
  - start with an empty dictionary.
    … and initial grapheme-to-phoneme (G2P) training data is generated manually
  - usually one single favorite G2P converter is used.
Semi-automatic pronunciation generation

- Manual pronunciation generation supported with estimated pronunciation
- Data-driven G2P converter for producing estimated pronunciation iteratively updated
1. Word-pronunciation pairs to train initial G2P converters.
2. A word is determined for pronunciation generation.
3. Pronunciations of that word are generated (by multiple G2P models).
4. Phoneme-level combination (PLC) of different G2P converter outputs.
Semi-automatic pronunciation generation

5. Display pronunciation to the user.
Semi-automatic pronunciation generation

6. User edits displayed pronunciation to the correct pronunciation.
Semi-automatic pronunciation generation

7. Word and corrected pronunciation are added to the dictionary.
8. After a certain number of words, the G2P converters update their grapheme-to-phoneme models.
Experimental Setup

- **Languages**
  - English, German, Spanish, Vietnamese, Swahili, Haitian Creole
  - Differing in their grade of G2P relationship
  - Reference dictionaries from CMU, GlobalPhone *(Schultz & Schlippe, 2014)*

- **G2P converters**
  - Sequitur G2P *(Bisani & Ney, 2008)*,
  - Phonetisaurus *(Novak et al., 2012)*,
  - Default&Refine *(Davel, 2005)*,
  - CART tree *(Lenzo, 1997)*

- **Phoneme-level combination (PLC)** *(Schlippe et al., 2014)*
Experimental Setup

- G2P model update interval (CPU time vs. G2P quality)
  - Fixed (Update after fixed number of new pronunciations)
  - Linear (linearly growing intervals)
  - Adaptive
    - First: after each word,
    - later: after certain number of predicted pronunciations needed corrections

- Our tradeoff: Logistic growth
  - Build G2P models frequently for small dictionaries
    ... where generation is cheap and gain is high
  - Stable G2P models for larger dictionaries are updated less often
    ... while still obeying an upper bound for updates
**Experimental Setup**

- Evaluated different sources for initial pronunciations
  - Pronunciations derived from 1:1 G2P mapping (most commonly associated phoneme for each grapheme)

<table>
<thead>
<tr>
<th>Word</th>
<th>enact</th>
<th>balls</th>
<th>moans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>EH N AE K T</td>
<td>B AO L ZH</td>
<td>M OWN ZH</td>
</tr>
<tr>
<td>Grapheme-based</td>
<td>en a c t</td>
<td>b a l l s</td>
<td>m o a n s</td>
</tr>
<tr>
<td>Grapheme-based (mapped)</td>
<td>EH N AX K T</td>
<td>B AX L L S</td>
<td>M AW AX N S</td>
</tr>
</tbody>
</table>

- Web-driven G2P converters / Web-derived Pronunciations (Wiktionary)
Evaluation Metric

- Human editing effort measured by
  - Total number of edits
    … not comparable between different dictionary sizes or languages
  - Cumulated phoneme error rate as a normed value:

\[
cPER(n) := \frac{\sum_{i=1}^{n} sub(w_i) + ins(w_i) + del(w_i)}{\sum_{i=1}^{n} phonemes(w_i)}
\]

… from beginning of editing process to current word \( w_i \)
… comparable across languages and dictionary sizes

- Human effort is analyzed in simulated experiments
  - … assuming that the developer changes the displayed phoneme sequence to the reference
Experiments: Single G2Pconverters

- Single G2P converters on 10k dictionaries (cPER)

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>vi</th>
<th>es</th>
<th>ht</th>
<th>sw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequitur</td>
<td>15.24</td>
<td>11.02</td>
<td>4.83</td>
<td>2.19</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>Phonetis.</td>
<td>15.28</td>
<td>11.10</td>
<td>4.42</td>
<td>2.28</td>
<td>0.43</td>
<td>0.21</td>
</tr>
<tr>
<td>D&amp;R</td>
<td>16.80</td>
<td>12.85</td>
<td>5.12</td>
<td>2.23</td>
<td>0.42</td>
<td>0.21</td>
</tr>
<tr>
<td>CART</td>
<td>20.01</td>
<td>13.89</td>
<td>5.20</td>
<td>2.56</td>
<td>0.36</td>
<td>0.26</td>
</tr>
</tbody>
</table>

- Languages with closer G2P relationship → less effort
- Sequitur performs best in three of six cases, closely followed by Phonetisaurus
- Phonetisaurus, Default&Refine and CART tree perform best in one case each
Experiments: G2P converter combination

- Phoneme-level combination (PLC)
  - cPER

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>es</th>
<th>vi</th>
<th>sw</th>
<th>ht</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best single G2P</td>
<td>15.24</td>
<td>11.02</td>
<td>2.19</td>
<td>4.42</td>
<td>0.21</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>PLC</td>
<td>14.47</td>
<td>10.81</td>
<td>2.00</td>
<td>3.78</td>
<td>0.13</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Relative to single</td>
<td>5.05$^s$</td>
<td>1.91$^s$</td>
<td>8.68$^s$</td>
<td>14.48$^s$</td>
<td>38.10$^s$</td>
<td>22.22$^s$</td>
<td>+15.07</td>
</tr>
</tbody>
</table>

- Always lower user editing effort
- On average: 15% relative reduction
- Between 1.9% and 38.2% relative depending on the language
Experiments: Initial Pronunciations

- Initial pronunciations
  - derived from 1:1 G2P mapping, Web-derived pronunciations (WDP)
  - included in phoneme level combination (PLC)
  - leave out after performance without them is better
    (cross-over based on average of all tested languages)

- 1:1 G2P mapping provides benefit in the beginning
- Wiktionary speeds up whole generation
Experiments: Initial Pronunciations

- Initial pronunciations
  - derived from 1:1 G2P mapping, Web-derived pronunciations (WDP)
  - included in phoneme level combination (PLC)

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>es</th>
<th>vi</th>
<th>sw</th>
<th>ht</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best single G2P</td>
<td>15.24</td>
<td>11.02</td>
<td>2.19</td>
<td>4.42</td>
<td>0.21</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>PLC</td>
<td>14.47</td>
<td>10.81</td>
<td>2.00</td>
<td>3.78</td>
<td>0.13</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Relative to single</td>
<td>5.05s</td>
<td>1.91s</td>
<td>8.68s</td>
<td>14.48s</td>
<td>38.10s</td>
<td>22.22s</td>
<td>+15.07</td>
</tr>
<tr>
<td>1:1 G2P mapping + PLC</td>
<td>14.45</td>
<td>10.81</td>
<td>1.97</td>
<td>3.73</td>
<td>0.12</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Relative to PLC</td>
<td>0.14</td>
<td>0.00</td>
<td>1.50</td>
<td>1.32s</td>
<td>7.69s</td>
<td>7.14s</td>
<td>+2.97</td>
</tr>
<tr>
<td>WDP + PLC</td>
<td>13.64</td>
<td>10.22</td>
<td>1.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative to PLC</td>
<td>5.74s</td>
<td>5.46s</td>
<td>6.50s</td>
<td></td>
<td></td>
<td></td>
<td>+5.90</td>
</tr>
</tbody>
</table>

- Still 6% reduction on top of PLC with WDP
- 1:1 G2P mapping only 3% on top of PLC
Conclusion and Future Work

- Semi-automatic pronunciation generation strategy
- Evaluated with languages with different G2P relationship
- 15% effort reduction with phoneme-level combination
- 6% on top with Web-driven G2P converters
- Cross-lingual G2P conversion (Schlippe et al., 2013) for initial pronunciations gave only slight improvement

Future work:
- Integrate additional G2P converters
- Improve phoneme-level combination
- Testing our semi-automatic pronunciation generation strategy with real human developers
Thank you!
References


References


