Features for Factored Language Models for Code-Switching Speech

Heike Adel, Katrin Kirchhoff, Dominic Telaar, Ngoc Thang Vu, Tim Schlippe, Tanja Schultz
Outline

- Introduction
  - Motivation
  - Seame Corpus
  - Main Contributions
- Factored Language Models
  - Features for Code-Switching Speech
- Experiments
- Conclusion
  - Summary
Motivation

- Code-Switching (CS) = speech with more than one language
- Exists in multilingual communities or among immigrants

Challenges:
- Multilingual models and CS training data necessary
SEAME corpus

- SEAME = South East Asia Mandarin-English
- Conversational speech, recorded from Singaporean and Malaysian speakers by [1]

Challenges
- much CS per utterance (∅: 2.6)
- short monolingual segments (mostly less than 1 sec, 2-4 words)
- not much training data for LM (575k words)

[1] Lyu, D.C. et al., 2010
Originally used: research project ‘Code-Switch‘ (NTU and KIT)
Main contributions

- Investigation of different features for Code-Switching speech

- Integration of factored language models into a dynamic one-pass decoder
Factored Language Models (FLMs) [2]

- Idea: word = feature bundle
  \[ w_t \equiv \{ f_t^1, f_t^2, \ldots, f_t^K \} \]

- Good e.g. in the case of
  - Rich morphology
  - Few training data
  => applicable to CS task

- Generalized backoff

Features: Words, POS, LID

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL dev</th>
<th>PPL eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (3-gram)</td>
<td>268.39</td>
<td>282.86</td>
</tr>
<tr>
<td>POS</td>
<td>260.70</td>
<td>267.86</td>
</tr>
<tr>
<td>LID</td>
<td>263.24</td>
<td>267.63</td>
</tr>
<tr>
<td>POS + LID</td>
<td>257.62</td>
<td>264.20</td>
</tr>
</tbody>
</table>

Problems:

- POS tagging of CS speech: challenging
- Accuracy of POS tagger: unknown

→ different clustering method may be more robust
Clusters based on word distributions in text [3]
- minimize average mutual information loss

Best number of classes in terms of PPL: 70

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL dev</th>
<th>PPL eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (3-gram)</td>
<td>268.39</td>
<td>282.86</td>
</tr>
<tr>
<td>POS + LID</td>
<td>257.62</td>
<td>264.20</td>
</tr>
<tr>
<td>Brown clusters</td>
<td>257.17</td>
<td>265.50</td>
</tr>
<tr>
<td>Brown clusters + POS</td>
<td>249.00</td>
<td>255.34</td>
</tr>
<tr>
<td>Brown cl + POS + LID</td>
<td>251.39</td>
<td>259.05</td>
</tr>
</tbody>
</table>

So far: clusters based on syntax or word distributions
- next step: semantic features

Features: Open Class Words

Definition: content words, e.g. nouns, verbs, adverbs

“open” because class can be extended with new words, e.g. “Bollywood”

→ open class words indicate semantic of sentence

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL dev</th>
<th>PPL eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 3-gram</td>
<td>268.39</td>
<td>282.86</td>
</tr>
<tr>
<td>Brown clusters + POS</td>
<td>249.00</td>
<td>255.34</td>
</tr>
<tr>
<td>Last oc word per speaker + Brown clusters + POS</td>
<td>247.18</td>
<td>252.37</td>
</tr>
</tbody>
</table>
Features: Open Class Word Clusters

- **Idea:**

  OC word 1, OC word 2, OC word 3, …, OC word 8, OC word 9

  Topic a
  - OC word 1
  - OC word 3
  - OC word 6

  Topic b
  - OC word 2
  - OC word 5
  - OC word 9

  Topic c
  - OC word 4
  - OC word 7
  - OC word 8

- Semantic clusters in comparison to distribution based clusters (oc Brown clusters)
Features: Semantic OC Word Clusters

- Clustering of open class word vectors
  - RNNLMs learn syntactic and semantic similarities [4]
  - RNNLMs represent words as vectors
  - apply clustering to these word vectors
    - k-means clustering
    - spectral clustering

Features: Semantic OC Word Clusters

- Experiments with different
  - Clustering methods
    - Brown, k-means, Spectral Clustering

- Monolingual and bilingual clusters
  - Monolingual Clusters
    - Based on English and Mandarin Gigaword data (2005)
  - Bilingual Clusters
    - Based on CS text
    - Mixed lines of Gigaword data

- Different numbers of clusters

- Lowest perplexity (247.24, but unclustered oc words: 247.18):
  - Spectral Clustering
  - Bilingual Clusters
  - 800 OC word clusters
FLMs: Decoding Experiments

Interpolation weight of FLM and n-gram

mixed error rate (on 20% of dev set)

FLM weight

0.2 0.4 0.5 0.55 0.6 0.65 0.7 0.8

43 42,5 42 41,5 41 40,5
## FLMs: Decoding Experiments (2)

### Decoding results

<table>
<thead>
<tr>
<th>Model</th>
<th>MER dev</th>
<th>MER eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 3-gram</td>
<td>39.96%</td>
<td>34.31%</td>
</tr>
<tr>
<td>POS</td>
<td>39.47%</td>
<td>33.46%</td>
</tr>
<tr>
<td>POS + LID</td>
<td>39.66%</td>
<td>33.30%</td>
</tr>
<tr>
<td>Brown clusters</td>
<td>39.45%</td>
<td>33.93%</td>
</tr>
<tr>
<td>Brown clusters + POS</td>
<td>39.30%</td>
<td>33.60%</td>
</tr>
<tr>
<td>Brown clusters + POS + LID</td>
<td>39.39%</td>
<td>33.16%</td>
</tr>
<tr>
<td>OC words + Brown clusters + POS</td>
<td>39.33%</td>
<td>33.15%</td>
</tr>
<tr>
<td>OC clusters + Brown clusters + POS</td>
<td>39.30%</td>
<td>33.16%</td>
</tr>
</tbody>
</table>
Conclusion

Summary

- Best features in terms of FLM perplexity: words + POS + Brown clusters + oc words
- Relative PPL reduction of up to 10.8% (eval)
- Best features in terms of MER: words + POS + Brown clusters (+ oc clusters)
- Relative MER reduction of up to 3.4% (eval)
THANK YOU FOR YOUR ATTENTION!
References


References


