A Generalized Method for Automated Multilingual Loanword Detection

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy

Introduction

• **Loanwords**: words incorporated from one language to another without translation
• If two words sound similar and have similar meanings, this is (usually) too coincidental to have occurred by chance
• We present a method to automatically detect loanwords between arbitrary language pairs
• Account for phonetic, semantic, orthographic, and articulatory features
• Evaluate on 12 language pairs, 4 unseen language pairs
• Our method achieves or exceeds SOTA and human performance
• Findings suggest features of loanwords allow generalization

Data

• **Wiktionary LoanWord (WikLoW) dataset**: 16 language pairs gathered from Wiktionary, with extensible method
• Positive loans augmented with:
  • **Synonyms** (similar meaning, different pronunciation)
  • **Hard Negatives** (different meaning, similar pronunciation)
  • **Randoms** (different meaning, different pronunciation)
• Converted to IPA using Epitran and articulatory features using PanPhon

Algorithm

• Extract 6 edit distances from PanPhon: **Fast Levenshtein Distance, Dolgo Prime Distance, Feature Edit Distance, Hamming Feature Distance, Weighted Feature Distance, Partial Hamming Feature Distance**
• Extract cosine similarity between word pairs from multilingual language models **MBERT and XLM-100**
• Deep neural network to score alignment between articulatory features
• **Binary classification**: Logistic Regressor, Neural Network, Support Vector Machine, Random Forest

Results

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>NN</th>
<th>SVM</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 (+)</td>
<td>85</td>
<td>86</td>
<td>84</td>
<td>85</td>
</tr>
</tbody>
</table>

Avg. positive F1 (%) of 4 classifiers

• Four experiments: Single Multilingual Model (SMM), pair-specific models, pruned training set (Small-SMM), unseen language pairs

Conclusion and Future Work

• We present an extensible method and novel baseline in loanword detection for arbitrary language pairs
• Automated loanword detection enables many downstream tasks
• Loanword knowledge is useful in, e.g., coreference resolution, NER, MT
• Parallel loanword corpora afford learning cross-lingual embedding mappings

Resources

• **Codebase**: https://github.com/csu-signal/loan-word-detection