Automatic Detection of Anglicisms for the Pronunciation Dictionary Generation: A Case Study on our German IT Corpus

Sebastian leidig, **Tim Schlippe**, Tanja Schultz
Motivation

From Microsoft's German website www.microsoft.de:

- “Zeigen Sie andere Apps für einfaches Multitasking neben dem Browser an.”
- “Internet Explorer nutzt Hardwarebeschleunigung. Websites werden schneller geladen, damit Sie noch reibungsloser surfen können.”
Motivation

With the globalization words from other languages come into a language without assimilation to the phonetic system of the new language.

To economically build up lexical resources with automatic or semi-automatic methods.

→ detect and treat them separately.
Overview

Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios

Input

word list
- word1
- word2
- word3
- word4
- word5
- word6

features

- grapheme perplexity
- g2p confidence
- hunspell lookup (native)
- hunspell lookup (English)
- Wiktionary lookup
- Google hit count

combination

- voting
- decision tree
- SVM

Output

classification
Outline

1. Motivation and Overview
2. Test Sets
3. Single Features
4. Combinations
5. Summary and Future Work

Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios
Test Sets - Domains

- **German IT website**
  - www.microsoft.de
  - 4.6k unique words

- **German general news**
  - www.spiegel.de
  - 6.6k unique words

- **Afrikaans**
  - NCHLT corpus *(Heerden, Davel, Barnard, 2013), (Basson, Davel, 2013)*
  - 9.4k unique words
Test Sets - Domains

- Tag for
  - “English”:
    - e.g. Software, Brain, …
  - Foreign hybrids
    - Compound words
      - e.g. Schadsoftware, …
    - Grammatically adapted words
      - e.g. downloaden, …

- Different word categories:
  - Abbreviations:
    - e.g. UV, CIA, …
  - Other foreign words
    - Compound words
      - e.g. Français, Niveau, …

- Decisions based on
  - Agreement of annotators
  - duden.de
Foreign words in different test sets

- **Microsoft-de**: 79% native, 15% English, 2% other foreign, 4% abbreviations
- **Spiegel-de**: 93% native, 2% English, 4% other foreign, 1% abbreviations
- **NCHLT-af**: 94% native, 2% English, 1% other foreign, 3% abbreviations
Single Features – Design Criteria

- Features trained on commonly available resources
  - Word lists, Pronunciation dictionaries, Spellchecker dictionaries, Wiktionary, Google

- Thresholds without supervised training
  - Comparison between English and native models

- New approaches
Grapheme Perplexity

Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios
Grapheme Perplexity

<table>
<thead>
<tr>
<th>Word</th>
<th>Perplexity(de)</th>
<th>Perplexity(en)</th>
<th>Perplexity diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entwicklungen</td>
<td>2.03</td>
<td>9.67</td>
<td>-7.64</td>
</tr>
<tr>
<td>Brain</td>
<td>19.88</td>
<td>5.58</td>
<td>14.30</td>
</tr>
</tbody>
</table>
Grapheme-to-Phoneme Confidence

Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios
## Grapheme-to-Phoneme Confidence

Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios

<table>
<thead>
<tr>
<th>Word</th>
<th>G2P conf-de</th>
<th>G2P conf (en)</th>
<th>G2P conf diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entwicklungen</td>
<td>9.90</td>
<td>35.72</td>
<td>-25.83</td>
</tr>
<tr>
<td>Brain</td>
<td>17.71</td>
<td>9.98</td>
<td>7.73</td>
</tr>
</tbody>
</table>

![Diagram showing grapheme-to-phoneme confidence](image)
Hunspell Lookup

word list
word1
word2
word3
word4

spellchecker dictionary
English: Hunspell-en

Hunspell

dictionary lookup
derive word forms

classification

2 features performed best

spellchecker dictionary
German: Hunspell-de

Hunspell

dictionary lookup
derive word forms

classification

word list
word1
word2
word3
word4
Hunspell Lookup

<table>
<thead>
<tr>
<th>Word</th>
<th>Stem</th>
<th>Dictionary(de)</th>
<th>Dictionary(en)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entwicklungen</td>
<td>Entwicklungs</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Brain</td>
<td>-</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

spellchecker dictionary

English: Hunspell-en

German: Hunspell-de

word list

word1

word2

word3

word4

classification

Hunspell

dictionary lookup

derive word forms

classification
## Wiktionary Lookup

- Check crowdsourced information from matrix language Wiktionary

<table>
<thead>
<tr>
<th>downloaden (Deutsch) [Bearbeiten]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb [Bearbeiten]</td>
</tr>
<tr>
<td>Anmerkung:</td>
</tr>
<tr>
<td>Der Gebrauch des Verbs schwankt. Es wird bevorzugt als Partikelverb mit abtrennbaren Verhpartikeln.</td>
</tr>
</tbody>
</table>

Herkunft:
Entlehnung aus dem Englischen to *download* → en „abladen, herunterladen“[2]
Google Hit Count

<table>
<thead>
<tr>
<th>Word</th>
<th>Results(de)</th>
<th>Results(en)</th>
<th>Normalized(de)</th>
<th>Normalized(en)</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entwicklungen</td>
<td>26.40M</td>
<td>0.58M</td>
<td>1.43 x 10^{-4}</td>
<td>1.87 x 10^{-7}</td>
<td>1.43 x 10^{-4}</td>
</tr>
<tr>
<td>Brain</td>
<td>18.00M</td>
<td>1940.00M</td>
<td>9.78 x 10^{-5}</td>
<td>6.22 x 10^{-4}</td>
<td>-5.24 x 10^{-4}</td>
</tr>
</tbody>
</table>

Result: Single Features

![Graph showing f-scores for different features]

- g2p confidence diff: 70%
- grapheme ppl diff: 67%
- Google hit count diff: 66%
- hunspell (English): 64%
- hunspell (native): 61%
- Wiktionary (native): 52%

- Microsoft-de: 41%
- Spiegel-de: 25%
- NCHLT-af: 26%
- hunspell (English): 36%
- hunspell (native): 27%
- Wiktionary (native): 32%

- Microsoft-de: 49%
- Spiegel-de: 39%
- NCHLT-af: 12%
- hunspell (English): 42%
- hunspell (native): 12%
- Wiktionary (native): 37%
Grapheme-to-Phoneme Confidence

Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios
Result: Single Features

On Spiegel-de test set: Higher ratio of words classified as English are wrong
Result: Combination

The diagram illustrates the f-score for different datasets and models:

- **Microsoft-de**:
  - Best single feature: 75%
  - Voting: 73%
  - Decision Tree: 74%
  - SVM: 70%

- **Spiegel-de**:
  - Best single feature: 62%
  - Voting: 62%
  - Decision Tree: 58%
  - SVM: 49%

- **NCHLT-af**:
  - Best single feature: 62%
  - Voting: 62%
  - Decision Tree: 32%
  - SVM: 54%
Challenges

- Performance after filtering difficult words (oracle)
Conclusion and Future Work

- Features based on available sources
- New approaches:
  - G2P confidence
  - Wiktionary
- Further features:
  - Part-of-speech (POS)
  - Context, trigger words
  - Capitalization
  - Translate and compare
благодарим за внимание!
References


References


