Motivation: Misspellings in Learner Texts

High error rate

Ich suche 3-4 Zimmer im Dreter Stock. [I am looking for 3-4 rooms on third floor.]  

[diagram showing phonetically similar misspellings with edit distance > 1]

Spellchecking tools are evaluated on different datasets and languages.

Existing tools mostly target errors of proficient writers.

The LeSpell Data Set

Uniform XML format for contextualized errors to allow for easy benchmarking experiments.

<table>
<thead>
<tr>
<th>Modality</th>
<th># Errors</th>
<th>Levensh.</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL-Spell (April 2018)</td>
<td>L2</td>
<td>6,251</td>
<td>1.25</td>
</tr>
<tr>
<td>MERLIN-CZ (Keiper et al., 2016)</td>
<td>L2</td>
<td>4,807</td>
<td>1.35</td>
</tr>
<tr>
<td>CITA (Barbagli et al., 2016)</td>
<td>L1</td>
<td>1,431</td>
<td>1.14</td>
</tr>
<tr>
<td>MERLIN-IT (Barbagli et al., 2016)</td>
<td>L2</td>
<td>2,137</td>
<td>1.18</td>
</tr>
<tr>
<td>Litkey (Quaranti et al., 2015)</td>
<td>L1</td>
<td>38,698</td>
<td>1.40</td>
</tr>
<tr>
<td>MERLIN-DE (Schneider-Allian et al., 2018)</td>
<td>L2</td>
<td>5,366</td>
<td>1.34</td>
</tr>
<tr>
<td>Skala (Schneider-Allian et al., 2018)</td>
<td>L1</td>
<td>423</td>
<td>1.25</td>
</tr>
</tbody>
</table>

LeSpell – A Multi-Lingual Benchmark Corpus of Spelling Errors to Develop Spellchecking Methods for Learner Language

Marie Bexte, Ronja Laarmann-Quante, Andrea Horbach, Torsten Zesch

Research Cluster “Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education”, FernUniversität in Hagen, Germany

DKPro Spelling

Language-agnostic: Requires dictionary, tokenizer, Named Entity Recognizer

Misspelling Detection

Tokenization

Tagging of Named Entities, Numbers, Punctuation

Comparison against Dictionary

Language Model

Error Correction

Custom Dictionary

Custom Weights

Custom LM

Misspelling Detection

TOEFL-Spell

METHOCAL-CZ

CITA

MERLIN-IT

Litkey

MERLIN-DE

Skala

Correction Recall@1: Comparison of Tools

TOEFL-Spell

MERLIN-CZ

CITA

MERLIN-IT

Litkey

MERLIN-DE

Skala

Correction Recall@1: Comparison of Methods

TOEFL-Spell

MERLIN-CZ

CITA

MERLIN-IT

Litkey

MERLIN-DE

Skala

Summary

LeSpell is a multi-lingual benchmark data set of language learner spelling errors in context

DKPro Spelling is a highly customizable spellchecking extension:

Default error correction setting (grapheme-based generation of 3 candidates, reranked Web1T trigrams) outperforms Hunspell & LanguageTool on 5 out of 7 LeSpell corpora

If we knew which method gave the correct candidate, results would improve

Grapheme-based correction works best, except for Skala

Upper bound not reached

Data & Code

This work was conducted in the framework of the Research Cluster DPU “Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education” of the FernUniversität in Hagen, Germany (https://doi.org/10.5281/zenodo.3929279)