Towards the New Czech Grammar-checker

RASLAN 2018

Vojtěch Mrkývka
mrkyvka@phil.muni.cz

December 7, 2018
Goal

- New grammar-checker of Czech
- Web-based application
- Using new and existing tools developed at MU
Motivation

- There are tools existing / in development at MU
- Current best Czech GC is part of proprietary system
- Create an alternative to applications like Grammarly but for Czech
The current interface

Rozhraní korektoru

Přišel jsem k golemovi. On se na mě otočil když jsem se nedíval.
The current interface

- Based on on-line text processor tinyMCE
- Mostly in JavaScript as tinyMCE modules
- Asynchronous process
- Communication with backend via AJAX
Processing diagram

tokenization

lemmatization & tagging

some module

correction displaying

some module

some module

some module
Correction displaying

The dog is runing.

Tokens to display mistake at: 6
Correction: 6/runing/running
# Implemented modules

<table>
<thead>
<tr>
<th>Correction</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>pre</th>
<th>rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misspellings (excl. proper nouns)</td>
<td>24</td>
<td>0</td>
<td>487</td>
<td>16</td>
<td>1,000</td>
<td>0.600</td>
</tr>
<tr>
<td>Misspellings (incl. proper nouns)</td>
<td>7</td>
<td>17</td>
<td>497</td>
<td>6</td>
<td>0,292</td>
<td>0.538</td>
</tr>
<tr>
<td>Vocalisation of prepositions</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Multiple whitespaces</td>
<td>4</td>
<td>0</td>
<td>515</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Whitespace in the interpunction proximity</td>
<td>7</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Conditionals</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Commas in a sentence</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1,000</td>
<td>0.429</td>
</tr>
</tbody>
</table>
A problem with testing

- Testing data were too small
- Mistakes were artificial

⇒ Need for API & collection of correctly annotated genuine texts
# Implemented modules

<table>
<thead>
<tr>
<th>Correction</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>pre</th>
<th>rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misspellings (excl. proper nouns)</td>
<td>24</td>
<td>0</td>
<td>487</td>
<td>16</td>
<td>1,000</td>
<td>0,600</td>
</tr>
<tr>
<td>Misspellings (incl. proper nouns)</td>
<td>7</td>
<td>17</td>
<td>497</td>
<td>6</td>
<td>0,292</td>
<td>0,538</td>
</tr>
<tr>
<td>Vocalisation of prepositions</td>
<td>4</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Multiple whitespaces</td>
<td>4</td>
<td>0</td>
<td>515</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Whitespace in the interpunction proximity</td>
<td>7</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Conditionals</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Commas in a sentence</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1,000</td>
<td>0,429</td>
</tr>
</tbody>
</table>
A problem with spell-checking

- Precision is low
- Not often updated dictionary
- ⇒ Method of adding new words, using different lexicon…
A problem with error reporting

- Allow users to flag miscorrections
- How to not display miscorrection afterwards?
- ⇒ Probably module-depending
Thank you for your attention!

This work was supported by the project of specific research Čeština v jednotě synchronie a diachronie (Czech language in unity of synchrony and diachrony; project no. MUNI/A/0862/2017).