Automatická kontrola textu pre slovenčinu

Ondrej Čičkán

Vedúci projektu: Ing. Marián Šimko PhD.

Automatická kontrola textu

- Detekcia chýb a ich oprava
- Typy chýb
  - Preklepy
  - Chýbajúca diakritika, interpunkcia
  - Gramatické chyby
  - Štylistické chyby

- Problémy slovenčiny:
  - Ohybné slovné druhy (inflexívnosť)
  - Voľné poradie slov vo vete
  - Veľa výnimiek

Metódy kontroly textu

- nezávislé od kontextu / závislé od kontextu
- Využívajúce:
  - Slovníky (Hunspell)
  - Pravidlá (LanguageTool)
  - Štatistiku (Korektor)

Test nástrojov na kontrolu slovenčiny

278 slov
34 chýb

Použitý nástroj / Typ chyby	Neexistujúce slová	Chybné slová		Čiarky	Štýl	Nezaradené	Celkom
		Preklepy	Gramatické				
Bez kontroly	20	5	2	3	2	2	**34**
Microsoft Word	4	5	2	3	1	1	**16**
LibreOffice	2	5	2	3	2	2	**16**
LanguageTool	3	5	2	3	1	1	**15**
Hunspell	3	5	2	3	2	2	**17**
Aspell	3	5	2	3	2	2	**17**
Chrome	3	5	2	3	2	2	**17**

Počet neopravených chýb

Korektor češtiny

- http://ufal.mff.cuni.cz/korektor
- Štatistický korektor

- Využíva kombináciu jazykových modelov založených na:
  - Slovných tvaroch
  - Leme
  - Gramatických kategóriách

- Model chýb

	Korektor	MS Word
1	91,6%	71,2%
2	97,2%	-
5	98,6%	-

	Korektor	MS Word
Precision	1,0	0,5
Recall	0,77	0,08
F-measure	0,87	0,14

Richter Michal, Straňák Pavel and Rosen Alexandr. *Korektor – A System for Contextual Spell-checking and Diacritics Completion* In Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), pages 1-12, Mumbai, India, 2012

- Open source
- Možnosť vytvoriť podporu pre ďalšie jazyky

Jazykové modely pre slovenčinu

- Textový korpus
- Vytvoriť jazykové modely
  - Pre rôzne črty jazyka (lema, gramatické kategórie)
  - Lematizátor
  - Nástroj na určovanie gramatických kategórií

- Nástroje na tvorbu jazykových modelov
  - SriLM
  - KenLM

Model chýb

- Church & Gale, 1991

- Textový korpus s chybami
  - Identifikovať chyby
  - Len chyby do editačnej vzdialenosti 1
  - Spočítať počet chýb pre každú operáciu

Testovanie

- Testovacie dáta:
  - Časť textového korpusu, ktorý nebol použitý na trénovanie
  - Diktát