Fixing grammar errors semi-automatically

Marcin Miłkowski & Daniel Naber
London, Wikimania 2014
LanguageTool WikiCheck

Overview

• A system that looks for style and grammar problems:
  – in Wikipedia articles
  – in changes to Wikipedia articles

• Used to correct > 2,000 errors in de.wikipedia.org in the last few months
  – works with other languages, too
Roadmap

- How to find style and grammar errors automatically
- How LanguageTool WikiCheck works internally
- How LanguageTool WikiCheck can help you and how you can help WikiCheck
- Q+A
About us

• Marcin Miłkowski
  – Associate professor at the Polish Academy of Sciences, Warsaw, Poland

• Daniel Naber
  – Software developer from Potsdam, Germany
  – Also: openthesaurus.de
How to Find Style and Grammar Errors

- Go to http://tools.wmflabs.org/languagetool
  - Demo: Check any Wikipedia page
  - Demo: Check the recent changes
Wikipedia Integration

Examples of False Alarms

• Word repetition:
  – “his great great grand son”

• Use simply: a Thousand:
  – “location=Thousand Oaks, Calif.”

• => Always check the suggestions of LanguageTool
Supported Languages

- Asturian, Belarusian, Breton, Catalan, Chinese, Danish, Dutch, English, Esperanto, French, Galician, German, Greek, Icelandic, Italian, Japanese, Khmer, Lithuanian, Malayalam, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Tagalog, Tamil (next version), Ukrainian

- Supported to a very different degree
- Not all languages have the 'Recent Changes' check activated. Talk to me to activate it.
How it Works Internally

- Based on LanguageTool
  [https://languagetool.org](https://languagetool.org)
- Open Source (LGPL)
- Written in Java
- Doesn't know grammar, but knows (typical) errors
  - The errors are found in the text by matching them to patterns
Error Pattern: Simple Example

• "Sorry for my bed English."
  – Spell checking won't help here

• Pattern for that error:

  <pattern>
    <token>bed</token>
    <token>English</token>
  </pattern>
Error Pattern Editor

- Demo of the Rule Editor: 'Sorry for my bed English' at http://community.languagetool.org/ruleEditor2/
Error Patterns: Some Numbers

• How many error patterns do we have?
  – About 1,000 for English (help needed!)
  – About 1,800 for German

• ...but only:
  – 97 for Spanish
  – 43 for Japanese
  – 26 for Swedish
  – For the full list, see https://languagetool.org/languages/
Error Patterns for Wikipedia

- For German, we have some Wikipedia-specific patterns:
  - “in letzter Zeit” (recently)
  - “Es wurde bewiesen, dass” (Studies show)

- See potential patterns for English at https://en.wikipedia.org/wiki/Weasel_word
Known Limitations

- Won't find all errors
- Sometimes complains about text which is correct
  - Sometimes get confused by Mediawiki syntax (solution: Parsoid? Help welcome!)
- "Watch this page" is always off by default
How WikiCheck can help you

...as a software developer

- Use our public HTTP API at http://wiki.languagetool.org/public-http-api
- Maybe help us integrate LanguageTool with the Visual Editor? http://wiki.languagetool.org/mediawiki-visual-editor
How WikiCheck can help you

• ...as a Wikipedia editor
  – Subscribe to the feed at http://tools.wmflabs.org/languagetool/feedMatches to get fresh errors from 'recent changes'
  – Send us feedback
Summary

- LanguageTool WikiCheck can already find many errors
- It finds errors by searching for error patterns
- Your help in writing error patterns, for any language, is very welcome
Thank you

- Contact: daniel.naber@languagetool.org
- http://tools.wmflabs.org/languagetool
- http://languagetool.org

License of these slides: Creative Commons Attribution Share-Alike Licence 3.0