Automatische Prüfung von Wikipedia-Artikeln

WikiCon 2013
23.11.2013

Daniel Naber

LT

Überblick

- Worum geht es hier (nicht)?
- Fehlerbeispiele
- Überblick LanguageTool
- LanguageTool Funktionsprinzip
- Fehlermuster
- LanguageTool und Wikipedia

Daniel Naber

- Entwickler von LanguageTool
  - Umfrage: Wer kennt LanguageTool schon?
- Entwickler von OpenThesaurus
  - Umfrage: Wer kennt OpenThesaurus?

Worum geht es hier **nicht**?

- Einfache Rechtschreibprüfung
- Prüfung MediaWiki-Syntax
- Prüfung von Textformatierungen (fett, kursiv etc.)
- Bot-Programmierung
  - LanguageTool nutzt die Wikipedia-API und die XML-Dumps

Worum geht es hier?

- Automatische Prüfung von:
  - Rechtschreibung - mit Beachtung des Kontext
  - Stil
  - Grammatik

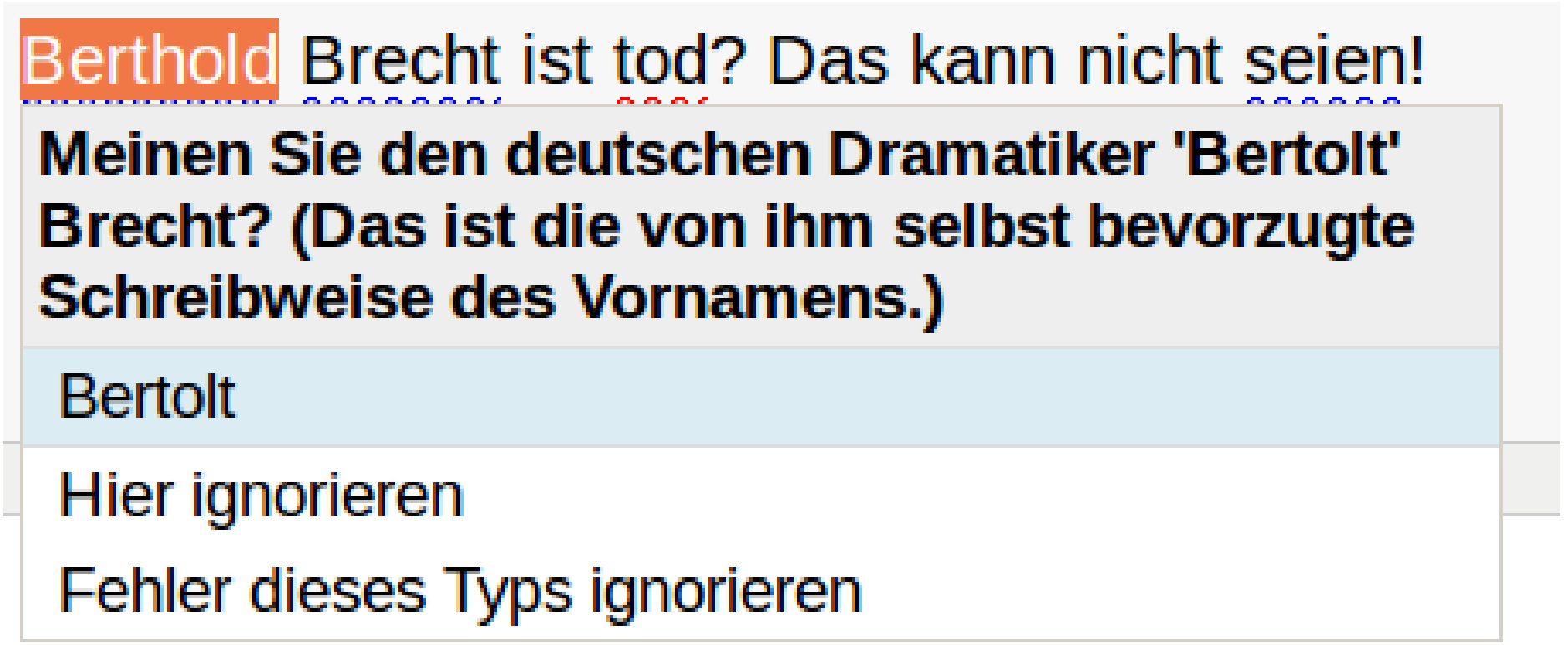
Grundidee

- Da anfangen, wo die klassische Rechtschreibprüfung aufhört
  - klassische Rechtschreibprüfung: prüft einzelne Wörter, ohne Kontext zu beachten

Finde die Fehler:

Berthold Brecht ist tod?
Das kann nicht seien!

Berthold Brecht ist tod? Das kann nicht seien!


Berthold Brecht ist tod? Das kann nicht seien!
Meinen Sie den deutschen Dramatiker 'Bertolt' Brecht? (Das ist die von ihm selbst bevorzugte Schreibweise des Vornamens.)
Bertolt
Hier ignorieren
Fehler dieses Typs ignorieren

LanguageTool: Überblick

- Open Source Stil- und Grammatikprüfung
- unterstützt 29 Sprachen, darunter Deutsch, Englisch, Französisch, Spanisch
- seit 2003
- Lizenz: LGPL
- nicht Wikipedia-spezifisch, aber gut für die Wikipedia nutzbar
- Homepage: http://languagetool.org

LanguageTool: Prinzip

- Text wird analysiert: Erkennung Satzgrenzen, Wortarten
  - „Hauses“: Nomen, Neutrum, Singular, Genitiv
- Text wird auf Fehlermuster durchsucht
  - LanguageTool kennt keine Grammatik, es kennt Fehler

Fehlermuster-Beispiel (vereinfacht)

```
<rule>
  <pattern>
    <marker>
      <token>Berthold</token>
    </marker>
    <token>Brecht</token>
  </pattern>
  <message>Meinen Sie den deutschen Dramatiker
     <suggestion>Bertolt</suggestion> Brecht?
  </message>
</rule>
```

Berthold Brecht ist tot.

Fehlermuster (1)

- Reguläre Ausdrücke:
  <token regexp="yes">Berth?old</token>
- Wortarten ansprechen, z.B. alle Verben:
  <token postag_regexp="yes"
  postag="VER:.*">
- Negation:
  <token negate="yes">Haus</token>
- ...und mehr: http://www.languagetool.org/development/

Fehlermuster (2)

- Deutsch: über 1700 Regeln als Fehlermuster
  - Englisch: 900 Regeln
  - Griechisch, Japanisch, Khmer: weniger als 30 Regeln

Fehlermuster: Grenzen

- Warum war „Brecht“ falsch im ersten Beispiel?

Berthold Brecht ist tod? Das kann nicht seien!

- Komplizierte Regeln in Java programmiert

Beispielfehler

- Seid gestern geht es mir besser.
- Das Ganze ist größer wie die Summe seiner Teile.
- Zertifiziert gemäß des aktuellen Standards.
- in Mühlheim an der Ruhr
- Sein Verhalten spiegelt seine Haltung wieder.
- Ich hatte eine Mittelohrenzündung.

Beispielfehler - mit Korrektur

- Seid gestern > Seit gestern
- größer wie > größer als
- gemäß des > gemäß dem
- Mühlheim an der Ruhr > Mülheim an der Ruhr
- spiegelt ... wieder > spiegelt ... wider
- Mittelohrenzündung =
  Mittel + Ohren + Zündung >
  Mittelohrentzündung

Wikipedia-spezifisches

- „seit Kurzem“ > relative Zeitangaben
- Wikipedia:Vermeide hohle Phrasen
  - „entgegen landläufiger Meinung“
  - „viele Wissenschaftler sagen“

LanguageTool und Wikipedia

- Fehlersammlung:
  http://community.languagetool.org/corpusMatch (aus dem Wikipedia-XML-Dump)
- Einzelne Seite prüfen:
  http://www.languagetool.org/wikicheck/
  auch mit Bookmarklet

Fehlermuster selber schreiben

- http://www.languagetool.org/ruleeditor/
  - prüft gegen 500.000 Sätze aus der Wikipedia
  - Advanced mode:
    Beispiel/Testsätze im XML

TODOs und Ideen

- Firefox-Plugin
  - kennt Mediawiki-Syntax nicht
  - nach Chrome portieren
- mehr Wikipedia-spezifische Regeln
- mehr Regeln (alle Sprachen)
- täglich alle Wikipedia-Änderungen prüfen?
- siehe http://wiki.languagetool.org/missing-features

Mitwirkende gesucht!

- Keine Programmierkenntnisse nötig
- Interesse?
  - mich hier ansprechen oder
  - auf der Mailing-Liste anmelden

http://languagetool.org

Zusammenfassung

- LanguageTool macht da weiter, wo die Rechtschreibprüfung aufhört
- LanguageTool hat mehrere Tools, um die Wikipedia zu prüfen
- Man kann ohne Programmierkenntnisse neue Fehlermuster schreiben
- **Mitwirkende gesucht!**

Vielen Dank!

Fragen?