Jasnopis: a new application for measuring readability of Polish texts

Włodzimierz Gruszczyński, Bartosz Broda,
Bartłomiej Nitoń, Maciej Ogrodniczuk

Warsaw School Of Social Sciences And Humanities
Institute of Computer Science, Polish Academy of Sciences

wgruszczynski@swps.edu.pl, bartosz.broda@gmail.com,
bartek.niton@gmail.com, maciej.ogrodniczuk@ipipan.waw.pl

Abstract

In the demo session we present a new application for automatic measuring of readability of Polish texts making use of two most common approaches to the topic: Gunning FOG index and Flesch-based Pisarek method and two novel methods: measuring distributional lexical similarity of a target text and comparing it to reference texts and using statistical language modeling for automation of a Taylor test.

Keywords: readability, text simplification, text understandability

Text readability is the measure used to determine how easy (or difficult) a given text can be to read and understand. In the demo session we present a new Web-based application for measuring the readability of a given text called Jasnopis (the name is a neologism consisting of words *jasno – clear* and *pisać – to write*).

At the moment, we focus on four methods of measuring readability:

1. FOG index (two variants: using words and base forms of words).
2. Pisarek index (four variants: linear and non-linear versions using words and base forms of words).
3. Automated Taylor test (two variants: based on perplexity and hit count).
4. Measuring similarity (two variants: based on binary features and *tf.idf* weighting method).

For of an automated version of Taylor test and similarity measuring we use the model trained on several well-known reference corpora:

- “Rzeczpospolita” corpus http://www.cs.put.poznan.pl/dweiss/rzeczpospolita),
- legal act corpus based on texts retrieved from the Internet System of Legal Acts,
- articles from “Wiedza i życie” archives,
- texts from Polish Wikipedia Corpus (http://clip.ipipan.waw.pl/PolishWikipediaCorpus)

as well as a newly created corpus of children’s literature.

The readability application has been implemented using the Django framework integrated with Celery task manager. It currently accepts three types of input sources: plain text, uploaded file and URL.