"Creation and Use of Social Media Resources"

Kaunas, 18-19 May 2017

The JANES Project: Tools and Resources for Linguistic Analysis and Automatic Processing of User-Generated Content in Slovene

Nikola Ljubešić
Jožef Stefan Institute, Ljubljana

JANES in a nutshell

- National basic research project
- 2014-2017
- 2 institutions, 8 team members
- Work packages
  - WP1 corpus construction
  - WP2 linguistic analysis
  - WP3 resource and tool development
- http://nl.ijs.si/janes

JANES corpus (v0.4)

Text types (size in tokens)

- Tweets 110M
- Fora 50M
- Blogs 30M
- News comments 15M
- Wikipedia talk pages 5M

Data processing

- normalization
- morphosyntactic annotation
- text- or user-level metadata enrichment

JANES resources and tools

TweetCat for Twitter data harvesting

- Especially useful for harvesting collections of low-frequency languages
- https://github.com/clarinsi/tweetcat

cSMTiser for text normalization

- Character-level SMT (Moses)
- State-of-the-art results (1st place in the CLIN27 shared task)
- Applied to Swiss German (baseline ~22%, cSMTiser ~90%), Slovene, Croatian, Serbian, Dutch; CMC, historical and dialectal texts
- https://github.com/clarinsi/csmtiser

JANES resources and tools

Janes-tagger for morphosyntactic annotation of non-standard text

- Model with ~95% accuracy on standard text achieves ~69% accuracy on non-standard user-generated content (900+ categories)
- Combining general and 75k of in-domain data, Brown clusters, normalization information yields accuracy of ~88%
- https://github.com/clarinsi/janes-tagger

Janes-Tag (http://hdl.handle.net/11356/1123) resource for learning normalization, tagging and named entity recognition, 75k tokens

ReLDI-NormTag-hr (http://hdl.handle.net/11356/1121) and **ReLDI-NormTag-sr** (http://hdl.handle.net/11356/1120) comparable resources for Croatian and Serbian

JANES resources and tools

Metadata enrichment

- Text-level enrichment
  - Technical (1-3) and linguistic (1-3) standardness (Ljubešić et al, 2015)
  - Sentiment (Fišer et al, 2016)
- User-level enrichment
  - Gender (Škrjanec et al, 2017)
  - Type (private, corporate) (Ljubešić et al, 2016)
  - Region (Čibej, 2016)

FRENK - annotation, analysis and identification of socially unacceptable discourse

- 2017-2020
- 4 partners - NLP and machine learning, linguistics, social sciences, law
- Facebook comments (mainstream media, Facebook pages)
- Annotation schema with 7 levels, professional annotators
- Predictions will be based on
  - text-level features
  - post-level features
  - discourse-level features
  - user-level features

The JANES Project: Tools and Resources for Linguistic Analysis and Automatic Processing of User-Generated Content in Slovene

Nikola Ljubešić
Jožef Stefan Institute, Ljubljana