Training data and tools for processing user-generated content in Slovene, Croatian and Serbian

Tomaž Erjavec[1], Nikola Ljubešić[1], Darja Fišer[2,1]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2]Department of Translation, Faculty of Arts, University of Ljubljana

SlaviCorp 2018, Prague
September 24 – 26, 2018

Introduction

- The language of social media (tweets, forums, blogs, etc.) is collectively known as user-generated content (UGC)
- In the last decade UGC has become a very studied and influential text type:
  - quantity, accessibility
  - knowledge and opinion
  - fake news, hate speech
- UGC often significantly differs from standard language: *Ali nisi bila včeraj na Bledu?* vs. *A nis bla včer na Bledu?*
- Common differences:
  - non-standard word spellings: *ne malo (nemalo)*
  - phonetic orthography: *jest, jst, js, ... (jaz)*
  - colloquial expressions: *oreng, orng (zelo)*
  - omissions of diacritics: *krizisce (križišče)*

Introduction

- Severe impact on automated text processing
  Gimpel et al. (2011): 97% PoS tagging accuracy on WSJ, but only 85% on a Twitter dataset, 5x increase in the error rate
- *Ali nisi bila včeraj na Bledu?* vs. *A* nis* bla* včer* na Bledu?*
- Two ways of overcoming this problem:
  - standardize the words in UGC, then use tools trained on standard language for further annotation;
  - train the tools with additional, UGC domain data, i.e. preform domain adaptation.

The topic of the talk

- Manually annotated datasets and tools that can be used to improve *automatic annotation of UGC text* for three South Slavic languages:
  - Slovene
  - Croatian
  - Serbian
- The datasets and tools cover six types of annotation:
  - tokenisation & sentence splitting,
  - word normalization,
  - morphosyntactic annotation & lemmatisation,
  - named entity recognition.

Data annotation workflow

- Annotation proceeded in a similar fashion for the 3 languages
- The Slovene datasets were produced first in the Janes project

Workflow

1. Annotation guidelines written
2. Student annotators trained on preliminary test data
3. Sampled data from the Janes corpus & imported to WebAnno
4. Files distributed to annotators; each file to 2 annotators
5. Disagreements checked by the curator
6. Output WebAnno files merged with their source TEI

- Annotation guidelines translated to English & annotation campaigns similar to Slovene performed for Croatian and Serbian
- This saved time and effort & produced harmonized resources

Slovene Datasets

Janes-Norm

- 150,000 words of Slovene UGC
- Manually corrected tokenisation, sentence segmentation and normalization of the words to standard Slovene
- Automatically assigned morphosyntactic descriptions (MSDs) and lemmas (on the basis of the standardized words)
- Technically, a difficult aspect is when one non-standard word is mapped to several standardised ones or vice versa.

Janes-Tag

- 55,000 words, a subset of Janes-Norm
- Manually corrected MSDs and lemmas
- V2 annotated also with named entities

Croatian and Serbian Datasets

- Croatian: ReLDI-NormTagNER-hr, 80,000 words
- Serbian: ReLDI-NormTagNER-sr, 80,000 words
- Manually annotated for all six annotation layers, same as Janes-Tag

All the datasets are available under CC licenses for download from the CLARIN.SI repository, as well as for exploration and analysis via its on-line concordancers KonText and noSketch Engine.

Annotation Tools

- We have also produced state-of-the-art annotation tools to enable non-standard language processing for the three languages and the six levels of annotation:
  - ReLDI tokeniser: tokenisation and sentence splitting
  - CSMTiser: word normalisation
  - JANES-tagger: MSD tagging and lemmatisation
  - JANES-NER: named entity recognition
- The tools are available from `https://github.com/clarinsi`.

ReLDI tokeniser

- Tokeniser and sentence segmenter, Python
- Manually specified rules & language-specific resources files, such as lists of abbreviations
- Two modes: standard language, non-standard language
- For non-standard language rules are less strict & there are a few additional rules describing phenomena typical for on-line communication, such as emoticons
- Evaluation on highly non-standard tweets:
  tokenisation 99.2%; sentence segmentation 86.3%

CSMTiser

- Performs word normalization
- Uses character-level statistical machine translation (Moses)
- Trained on:
  - Slovenian: Janes-Norm,
  - Croatian: ReLDI-NormTagNER-hr,
  - Serbian: ReLDI-NormTagNER-sr
- For Slovene non-standard Twitter data:
  character-level accuracy = 98.5%, (non-normalized = 94.8%)

JANES-tagger

- Performs morphosyntactic tagging and lemmatisation
- Based on Conditional Random Fields
- Trained on standard-language datasets for three languages supplemented with Janes-Tag, ReLDI-NormTagNER-hr, and ReLDI-NormTagNER-sr
- Token-level accuracy on Slovene Twitter:
  - standard tagger = 69%
  - adapted tagger = 86%
  - (on standard data = 94%)

JANES-NER

- Performs Named Entity Recognition
- Based on Conditional Random Fields
- Trained on the same three datasets as the JANES-tagger
- F1 of JANES-NER:
  - overall: 0.69
  - person: 0.92
  - location: 0.80
  - organisation: 0.56
  - other class: 0.30

Conclusions

- Presented manually annotated datasets and tools for normalisation, morphosyntactic tagging, lemmatisation and named entity recognition of non-standard Slovene, Croatian, and Serbian.
- All datasets & tools freely available

Further work:

- Making the tools better: from CRF and SMT to neural networks

Training data and tools for processing user-generated content in Slovene, Croatian and Serbian

Tomaž Erjavec[1], Nikola Ljubešić[1], Darja Fišer[2,1]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2]Department of Translation, Faculty of Arts, University of Ljubljana

SlaviCorp 2018, Prague
September 24 – 26, 2018