hr500k – A Reference Training Corpus of Croatian

Nikola Ljubešić[1], Željko Agić[2], Filip Klubička[3],
Vuk Batanović[4], Tomaž Erjavec[1]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2]IT University of Copenhagen
[3]ADAPT Centre, Dublin Institute of Technology
[4]School of Electrical Engineering, University of Belgrade

JTDH 2018, Ljubljana
September 20 – 21, 2018

Introduction

- Manually annotated datasets – crucial for today’s NLP, which is based on supervised machine learning
- Until recently, Croatian was heavily under-resourced
- We present the central resource for building NLP tools for Croatian: a multi-annotated corpus of 500 thousand tokens
- hr500k is annotated on the levels of:
  1. document, sentence and token segmentation,
  2. morphosyntax,
  3. lemmas,
  4. dependency syntax,
  5. semantic roles and
  6. named entities.
- Two main parts of the talk:
  - Contents of the corpus
  - The story of building the corpus

Corpus in numbers

Documents	900
Sentences	24,794
Tokens	506,457
Types	73,548
Lemmas	34,329
MSDs	768

articles	blogs	forums	other
57.63%	20.6%	14.64%	7.13%

general	51.89%	education	3.61%
music	8.43%	religion	2.87%
medicine	7.63%	sports	2.74%
business	6.93%	listings	2.42%
tech	6.92%	culture	1.97%
lifestyle	4.59%		

Morphosyntax and lemmas

Morphosyntax

- Multiple annotation and consolidation rounds using the MULTEXT-East V5 morphosyntactic specifications
- Automatic mapping to UPOS, except for abbreviations

Lemmas

- Harmonisation with the inflectional lexicon hrLex

Way forward

- How to enable third parties to improve the hr500k and hrLex resources?
- How to cover low-frequency phenomena / improve tagger accuracy on these phenomena?

Dependency and shallow semantic parsing

Dependency syntax

- 197,028 tokens (2/5 of the corpus) annotated with Universal Dependencies v2.0
- Old formalism (PDT-based) also kept in the corpus
- Way forward: documentation and consolidation of annotations with Serbian and Slovene UD (any volunteers?)

Semantic role labeling

- Results of a bilateral Croatian-Slovene project
- 83,630 tokens annotated with a PDT-like formalism
- Single annotator, complex cases resolved by the consortium

Named entity recognition

Annotation procedure

- Annotation guidelines taken from the Slovene ssj500k corpus (hr500k's older cousin)
- WebAnno annotation by two annotators, conflict resolution by a third annotator
- PER (29%), DERIV-PER (1%), LOC (27%), ORG (27%), MISC (15%)
- Regardless of the strict guidelines and double annotation, some imperfections on the global level still remain

Way forward

- Global lexical consolidation planned for the near future
- Coreference resolution? (in the works for Serbian)

Corpus availability

Download

http://hdl.handle.net/11356/1183

Search

KonText:
https://www.clarin.si/kontext/first_form?corpname=hr500k

noSketch Engine:
https://www.clarin.si/noske/run.cgi/corp_info?corpname=hr500k

Models

Annotation tools for Croatian at:
https://github.com/clarinsi/

History of the resource

SETimes.HR

- 2012 - no freely available tagger for Croatian, no training data to train a tagger
- Željko and Nikola said – enough is enough, let us build a training corpus from the SETimes newspaper parallel dataset
- 87 thousand tokens in size
- Added initial syntactic and named entity annotations

SETimes.HR+

- Corpus extended with
  - Newspaper datasets collected for a student project in named entity recognition
  - Sample of sentences selected and annotated via a student-based crowd-sourcing campaign
- Improved variability of the resource

History of the resource (contd.)

hr500k

- Extension with additional 320k tokens
- In two phases:
  1. Analyze the current content for genre and topic distribution
  2. extend the resource to be representative of Croatian language? (i.e. the Croatian texts to be processed with the tools trained on this dataset)
- Manual selection of documents from the hrWaC web corpus to obtain a genre- and topic-representative dataset
- Harmonized morphosyntactic and lemma level
- Added new syntax, semantic role and named entity layers

Conclusion

- Example of a bottom-up approach to building resources[1] with (almost) no resources[2]
- Eventually became a large resource with many annotation layers (of varying quality, for now)
- No reason not to do it for other languages (SETimes.SR talk in the afternoon)
- Ways forward:
  - Harmonization of specific annotation layers inside the resource, but also across resources / languages (UD!)
  - Adding additional annotation layers (verbal MWEs - two different annotation layers!)
  - Size will be kept fixed for some time
- Many open questions, primarily on collaborative efforts in improving specific layers

hr500k – A Reference Training Corpus of Croatian

Nikola Ljubešić[1], Željko Agić[2], Filip Klubička[3],
Vuk Batanović[4], Tomaž Erjavec[1]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[2]IT University of Copenhagen
[3]ADAPT Centre, Dublin Institute of Technology
[4]School of Electrical Engineering, University of Belgrade

JTDH 2018, Ljubljana
September 20 – 21, 2018