Text processing tools for Latvian

Roberts Darģis, Lauma Pretkalniņa, Inguna Skadiņa
University of Latvia, IMCS, AILab

Outline of the presentation

- Facts about Latvian language
- Corpora and other resources
- List of text processing tools
- More detailed description of tools
- Language independent neural network morphological tagger

The Latvian language

- Latvian is an inflective language with rather free word order.
- Latvian features highly ambiguous set of morphological markers,
  - e.g., nouns in Latvian have 29 graphically different endings and 13 of them are unambiguous
- Nominals are inflected 5-7 cases, 2 numbers, and 2 genders
- Verbs have 5 moods, 3 persons, 6 tenses, 2 numbers, 2 voices and direct/reflexive distinction

Corpora and other resources

- Annotated corpora:
  - Morphology (111K tokens)
  - Syntax (Latvian Treebank / UD, 56K tokens)
  - NER (~150K tokens)
  - Valence samples for popular verbs
- Word embeddings
- Wordlists
  - Tēzaurs (dictionary, morphology, synonym pairs)
  - Place and person names
- Various unannotated texts
  - Balanced – 5,5M tokens
  - Blogs, parliament speeches, Wikipedia, etc.

Text processing tools I

- Tokenization, sentence splitting
  https://github.com/PeterisP/morphology
- Morphological analyzer
  https://github.com/PeterisP/morphology
- Morphological tagger (CMM)
  https://github.com/PeterisP/LVTagger
- NER tagger (CRF)
  https://github.com/PeterisP/LVTagger

Test processing tools II

- State of art morphological tagger for Latvian
  - Neural network based https://github.com/PeterisP/tf-morphotagger
- Normalizers for crippled text
  - For historical texts, can be recustomized for other uses https://github.com/LUMII-AILab/Transliterator
  - For web texts https://bitbucket.org/Ginta/ruukjiishi
- Text segmentation tool
  - Intended for domain name analysis, https://github.com/lauma/LVSegmenter
- UD based experimental syntactic parser

Tokenization

Rule-based tokenizer, containing definitions :

- Initial (A. Bērziņš = “A.” “Bērziņš”)
- Time (“12:34”, “12:54:32”) and
- Date (“2015.12.12”, “2015-12-12”)
- Numbers
  - Common numbers with thousand separator (space or apostrophe) and decimal separator (dot or comma).
  - Fractions (“54/100”)
  - Ordinal numbers in Latvian (ends with dot “1.”, “2016.”)
- E-mail, URL
- Repetitive punctuation (“!?!?!”, “….”)
- Common abbreviations and multiword conjunction (“piem.”, “u.c.”, “p.k.”, “it kā”, “gan arī”, “droši vien”)

Sentence splitting

- Based on tokenization.
- Sentence is split if at least one of following conditions are met:
  - Token consists only of end marks or their combinations (period, question mark, exclamation mark)
  - Sentence length capacity is reached (default 50 tokens)
  - End of the line (or document)

Morphological analyzer

- Single-token scope, gives all possible lemmas and feature sets
- Used to generate possible analysis variants
- Can also return lemma for given token and features (useful for DNN tagger)
- MULTEXT-East based tagset

Analysis of word “roku”

"Lemma":"roka",
"Part of speech":"Noun",
"Noun type":"Common noun",
"Gender":"Feminine",
"Case":"Accusative",
"Number":"Singular",
"Declension":"4"

"Lemma":"roka",
"Part of speech":"Noun",
"Noun type":"Common noun",
"Gender":"Feminine",
"Case":"Genitive",
"Number":"Plural",
"Declension":"4"

"Lemma":"rakt",
"Part of speech":"Verb",
"Tense":"Present",
"Mood":"Indicative",
"Number":"Singular",
"Conjugation":"1",
"Reflexive":"No",
"Person":"1",
"Voice":"Active"

CMM morphological tagger

- Selects the best feature set from variants provided by the analyzer
- Based on conditional Markov model (CMM)
- POS error rate 4.9%
- Full morphological tag error rate 8.6%

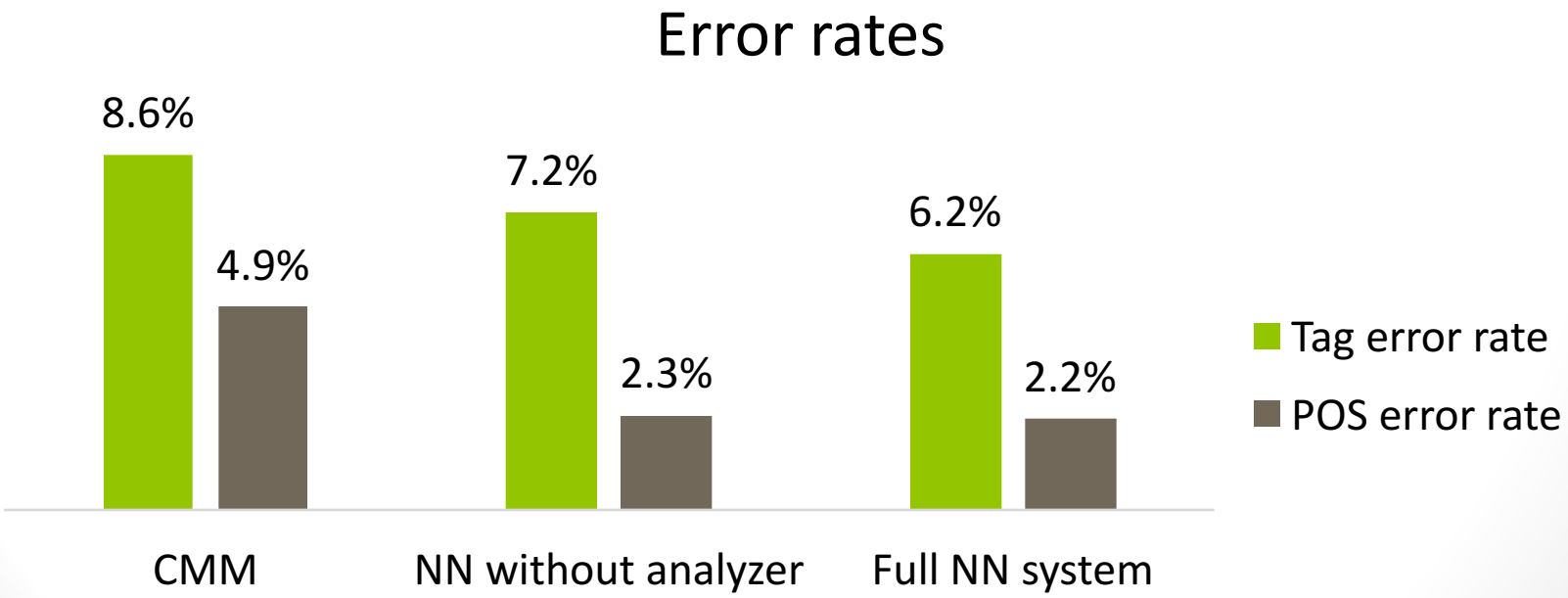
NER Tagger

- Based on Stanford NER system with extended feature set and an extensive gazetteer

Entity type	F1	P	R
location	86.9	84.2	89.9
media	77.2	95.1	65.0
organization	74.0	77.5	70.9
person	86.8	89.1	84.6
product	14.0	39.3	8.5
sum	94.1	97.3	91.2
time	88.3	92.7	84.4
Total	84.6	91.0	79.1

Neural network based morphological tagger

- Recently developed state of art morphological and Part of Speech tagger for Latvian
- Written in Python using TensorFlow



Tools and data required for NN morphological tagger

- Text and sentence tokenization
- Word embeddings (or large unannotated text)
- Morphologically annotated corpus
  - for most EU languages UD corpus can be used
  - unless something better is available

Relevant publications

- Paikens, P. (2016). Deep Neural Learning Approaches for Latvian Morphological Tagging. *Proceedings of the 7th International Conference: Human Language Technologies – The Baltic Perspective (Baltic HLT 2016).*
- Znotiņš, A. and Paikens, P. (2014). Coreference resolution for Latvian. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation.*
- Paikens, P., Rituma, L., and Pretkalniņa, L. (2013). Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) NEALT Proceedings Series 16,* pages 267–278, Oslo.
- Pretkalniņa L., Paikens P., Grūzītis N., Rituma L., Spektors A. Making Historical Latvian Texts More Intelligible to Contemporary Readers. Proc. of LREC 2012 Workshop "Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects", Istanbul, Turkey, 2012, pp. 29–35
- Paikens, P. (2007). Lexicon-based morphological analysis of Latvian language. *Proceedings of 3rd Baltic Conference on Human Language Technologies (HLT 2007).* (SCOPUS)

Thank you!

. . . ? . . .

Mākslīgā intelekta laboratorija
LU MII

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA