Open-Source Taggers for (Czech) POS Tagging and NE Recognition

Jana Straková, Milan Straka, Jan Hajič

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{strakova,straka,hajic}@ufal.mff.cuni.cz

10^{th} March 2014

Outline

- `http://ufal.mff.cuni.cz/morphodita`
  MorphoDiTa – morphologic dictionary and tagger
- `http://ufal.mff.cuni.cz/nametag`
  NameTag – NE recognizer
- `http://ufal.mff.cuni.cz/cnec`
  CNEC 2.0 – Czech Named Entity Corpus 2.0
  (joint work of Magda Ševčíková, Zdeněk Žabokrtský, Jana Straková and Milan Straka).

Outline

- `http://ufal.mff.cuni.cz/morphodita`
  MorphoDiTa – morphologic dictionary and tagger
- `http://ufal.mff.cuni.cz/nametag`
  NameTag – NE recognizer
- `http://ufal.mff.cuni.cz/cnec`
  CNEC 2.0 – Czech Named Entity Corpus 2.0
  (joint work of Magda Ševčíková, Zdeněk Žabokrtský, Jana Straková and Milan Straka).

Outline

- `http://ufal.mff.cuni.cz/morphodita`
  MorphoDiTa – morphologic dictionary and tagger
- `http://ufal.mff.cuni.cz/nametag`
  NameTag – NE recognizer
- `http://ufal.mff.cuni.cz/cnec`
  CNEC 2.0 – Czech Named Entity Corpus 2.0
  (joint work of Magda Ševčíková, Zdeněk Žabokrtský, Jana Straková and Milan Straka).

Motivation

POS tagger and NE recognizer which would:

- provide state-of-the-art results for Czech,
- be well suited and trainable for rich morphology languages,
- be distributed along with trained models,
- allow the user to train custom models,
- be extremely efficient in RAM and disc usage,
- offer API in multiple programming languages,
- be open-source, free software.

MorphoDiTa: Morphologic Dictionary and Tagger

MorphoDiTa performs:

- tokenization,
- morphologic analysis,
- lemmatization,
- morphologic generation,
- tagging.

Czech morphology:

based on Jan Hajič's Morfflex CZ

`http://ufal.mff.cuni.cz/morfflex`

MorphoDiTa: Morphologic Dictionary and Tagger

MorphoDiTa performs:

- tokenization,
- morphologic analysis,
- lemmatization,
- morphologic generation,
- tagging.

Czech morphology:

based on Jan Hajič's Morfflex CZ

`http://ufal.mff.cuni.cz/morfflex`

MorphoDiTa Demo

`http://ufal.mff.cuni.cz/morphodita/demo`

Morphologic Dictionary

Inflective languages with large number of endings:

- “zelený” (“green” in Czech): “zelený”, “zelenější, zelenému, etc. – several tens of forms for this type of adjectives
- 168K unique forms and 72K lemmas in a 2M corpus (PDT 2.5, Bejček et al., 2012)
- It is crucial to handle the endings effectively to reduce processing costs.

Templates

MorphoDiTa creates a set of “templates” from a given resource with lemmas, forms and tags (without language knowledge).

Morphologic Dictionary

Inflective languages with large number of endings:

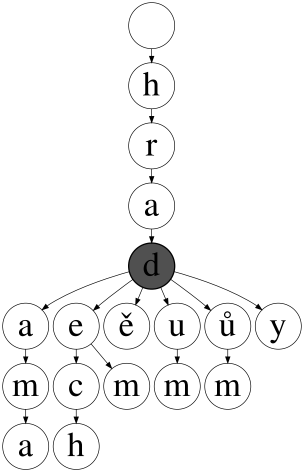
- “zelený” (“green” in Czech): “zelený”, “zelenější, zelenému, etc. – several tens of forms for this type of adjectives
- 168K unique forms and 72K lemmas in a 2M corpus (PDT 2.5, Bejček et al., 2012)
- It is crucial to handle the endings effectively to reduce processing costs.

Templates

MorphoDiTa creates a set of “templates” from a given resource with lemmas, forms and tags (without language knowledge).

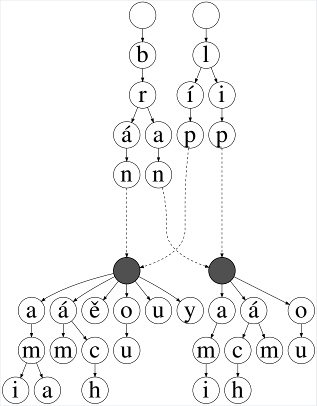
Morphologic Dictionary

Template set creation



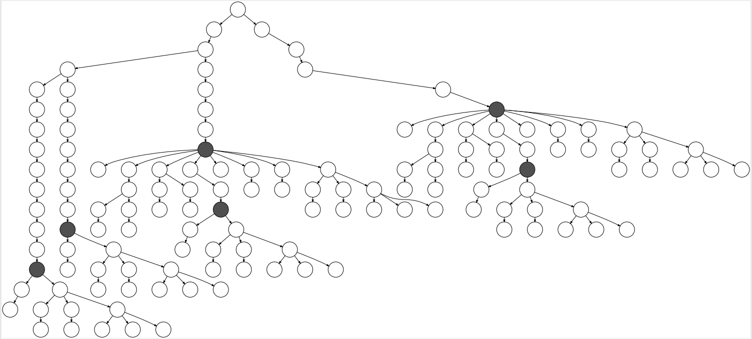
Morphologic Dictionary

Template set creation – cont.



Morphologic Dictionary

Template set creation – cont.



Morphologic Dictionary

Template set creation – cont.

- Czech Morfflex contains 120M form-tag, 1M unique lemmas, 3992 tags; total size $6.7 \cdot 10^9$ bytes.
- 7080 templates created
- whole dictionary encoded using 3M template instances
- binary form of the dictionary uses 2MB (3000 smaller)

POS Tagger Methodology

- for each form, morphologic dictionary suggests all lemma-tag candidates
- for each sentence, candidates are disambiguated by tagger
- tagger based on (Spoustová et al., 2009)
- supervised, rich feature averaged perceptron (Collins, 2002)

POS Tagger Evaluation

Tagger	Task	Accuracy
Morče	tag	95.67%
Featurama	tag	95.66%
MorphoDiTa	tag	95.75%
MorphoDiTa	lemma	97.80%
MorphoDiTa	lemma+tag	95.03%
MorphoDiTa	tag-first two pos.	99.18%

Table : Evaluation of Czech POS taggers.

POS Tagger Performance

Task	System	Words/s	RAM	Model size
tag	Morče	1K	902MB	178MB
tag	Featurama	2K	747MB	210MB
tag	MorphoDiTa	10K	52MB	16MB
first two pos.	MorphoDiTa	200K	15MB	2MB

Table : Evaluation of the POS tagger throughput, RAM and model size on a standard desktop computer.

NameTag

- tool for named entity recognition (NER),
- identifies and classifies named entities (proper names) in text,
- distributed with trained models,
- allows custom model training.

NameTag Demo

http://ufal.mff.cuni.cz/nametag/demo

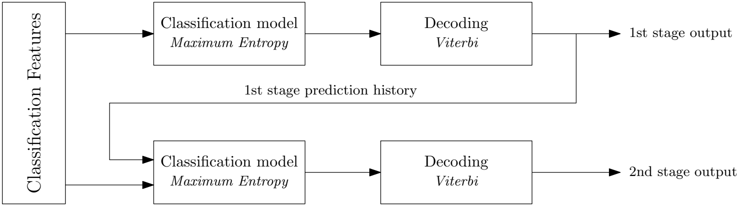
NER Methodology

- based on (Straková et al., 2013)
- supervised machine learning
- for each word (token), classification features are extracted:
    - morphological analysis (MorphoDiTa),
    - two-stage prediction (Ratinov and Roth, 2009)
    - gazetteers
    - word clustering (Brown, 1992).

NER Methodology

- based on (Straková et al., 2013)
- supervised machine learning
- for each word (token), classification features are extracted:
  - morphological analysis (MorphoDiTa),
  - two-stage prediction (Ratinov and Roth, 2009)
  - gazetteers
  - word clustering (Brown, 1992).

NER Methodology



- ME model predicts for each word in a sentence, a full probability distribution of its classes and position with respect to an entity,
- global optimization via dynamic programming determines the optimal combination of classes and NE chunks within sentence.

NER Evaluation

System	F_1 (42 classes)	F_1 (7 classes)
(Ševčíková et al., 2007)	62.00	68.00
(Kravalová and Žabokrtský, 2009)	68.00	71.00
(Konkol and Konopík, 2013)	NA	79.00
(Straková et al., 2013)	79.23	82.82
NameTag CNEC 1.1	77.88	81.01
NameTag CNEC 2.0	77.22	80.30

Table : Evaluation of the Czech NE recognizers.

NER Performance

Corpus	Words/s	RAM	Model size
CNEC 1.1	40K	54MB	3MB
CNEC 2.0	45K	65MB	4MB

Table : Evaluation of the NE recognizer tagger throughput, RAM and model size on a standard desktop computer.

Release

MorphoDiTa and NameTag available as

- standalone tool (precompiled)
- C++ source code with Python and Java bindings
- web service
- on-line demo

License

- source code under LGPL
- models under CC BY-NC-SA

Release

MorphoDiTa and NameTag available as

- standalone tool (precompiled)
- C++ source code with Python and Java bindings
- web service
- on-line demo

License

- source code under LGPL
- models under CC BY-NC-SA

CNEC 1.0

CNEC 1.0

- corpus of Czech named entities
- entities classified into two-level hierarchy
- 42 fine-grained classes, 7 coarse classes
- ambiguous labeling allowed
- embedding entities allowed

CNEC 1.1

fixes issues of CNEC 1.0

CNEC 2.0

- 8993 Czech sentences with 35220 NEs
- classification hierarchy changed (!)
- two-level hierarchy of 46 NEs
- released under CC BY-NC-SA
- `http://ufal.mff.cuni.cz/cnec`
- for list of changes, see documentation:
  `http://ufal.mff.cuni.cz/cnec/cnec2.0`

Conclusion

- http://ufal.mff.cuni.cz/morphodita
- http://ufal.mff.cuni.cz/nametag
- http://ufal.mff.cuni.cz/cnec