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

10th 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).
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 performs:

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

Czech morphology:

based on Jan Hajič’s Morfflex CZ

http://ufal.mff.cuni.cz/morfflex
http://ufal.mff.cuni.cz/morphodita/demo
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).
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).
Template set creation

Introduction
MorphoDiTa
NameTag
Release
CNEC 2.0
Conclusion
Jana Straková, Milan Straka, Jan Hajič
Open-Source Taggers for POS Tagging and NER
Morphologic Dictionary

Template set creation – cont.

Introduction

MorphoDiTa

NameTag

Release

CNEC 2.0

Conclusion

Jana Straková, Milan Straka, Jan Hajič

Open-Source Taggers for POS Tagging and NER
Template set creation – cont.
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)
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)
### Table: Evaluation of Czech POS taggers.

<table>
<thead>
<tr>
<th>Tagger</th>
<th>Task</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morče</td>
<td>tag</td>
<td>95.67%</td>
</tr>
<tr>
<td>Featurama</td>
<td>tag</td>
<td>95.66%</td>
</tr>
<tr>
<td>MorphoDiTa</td>
<td>tag</td>
<td>95.75%</td>
</tr>
<tr>
<td>MorphoDiTa</td>
<td>lemma</td>
<td>97.80%</td>
</tr>
<tr>
<td>MorphoDiTa</td>
<td>lemma+tag</td>
<td>95.03%</td>
</tr>
<tr>
<td>MorphoDiTa</td>
<td>tag-first two pos.</td>
<td>99.18%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Task</th>
<th>System</th>
<th>Words/s</th>
<th>RAM</th>
<th>Model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>tag</td>
<td>Morče</td>
<td>1K</td>
<td>902MB</td>
<td>178MB</td>
</tr>
<tr>
<td>tag</td>
<td>Featurama</td>
<td>2K</td>
<td>747MB</td>
<td>210MB</td>
</tr>
<tr>
<td>tag</td>
<td>MorphoDiTa</td>
<td>10K</td>
<td>52MB</td>
<td>16MB</td>
</tr>
<tr>
<td>first two pos.</td>
<td>MorphoDiTa</td>
<td>200K</td>
<td>15MB</td>
<td>2MB</td>
</tr>
</tbody>
</table>

**Introduction**

MorphoDiTa

**NameTag**

**Release**

CNEC 2.0

**Conclusion**

Jana Straková, Milan Straka, Jan Hajič

Open-Source Taggers for POS Tagging and NER
tool for named entity recognition (NER),
identifies and classifies named entities (proper names) in text,
distributed with trained models,
allows custom model training.
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).
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).
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.
Table: Evaluation of the Czech NE recognizers.

<table>
<thead>
<tr>
<th>System</th>
<th>$F_1$ (42 classes)</th>
<th>$F_1$ (7 classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Ševčíková et al., 2007)</td>
<td>62.00</td>
<td>68.00</td>
</tr>
<tr>
<td>(Kravalová and Žabokrtský, 2009)</td>
<td>68.00</td>
<td>71.00</td>
</tr>
<tr>
<td>(Konkol and Konopík, 2013)</td>
<td>NA</td>
<td>79.00</td>
</tr>
<tr>
<td>(Straková et al., 2013)</td>
<td>79.23</td>
<td>82.82</td>
</tr>
<tr>
<td>NameTag CNEC 1.1</td>
<td>77.88</td>
<td>81.01</td>
</tr>
<tr>
<td>NameTag CNEC 2.0</td>
<td>77.22</td>
<td>80.30</td>
</tr>
</tbody>
</table>
### Evaluation of the NE recognizer tagger throughput, RAM and model size on a standard desktop computer.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Words/s</th>
<th>RAM</th>
<th>Model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNEC 1.1</td>
<td>40K</td>
<td>54MB</td>
<td>3MB</td>
</tr>
<tr>
<td>CNEC 2.0</td>
<td>45K</td>
<td>65MB</td>
<td>4MB</td>
</tr>
</tbody>
</table>

**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

- 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
- for list of changes, see documentation: [http://ufal.mff.cuni.cz/cnec/cnec2.0](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