Improving Morphosyntactic Tagging of Slovene by Tagger Combination

Jan Rupnik
Miha Grčar
Tomaž Erjavec

Jožef Stefan Institute
Outline

• Introduction
• Motivation
• Tagger combination
• Experiments
POS tagging

Part Of Speech (POS) tagging: assigning morphosyntactic categories to words

Veža je smrdela po kuhanem zelju in starih, cunjastih predpražnikih.
Slovenian POS

- multilingual MULTEXT-East specification
- almost 2,000 tags (morphosyntactic descriptions, MSDs) for Slovene
- Tags: positionally coded attributes
- Example: MSD Agufpa
  - Category = Adjective
  - Type = general
  - Degree = undefined
  - Gender = feminine
  - Number = plural
  - Case = accusative
State of the art: Two taggers

- Amebis d.o.o. proprietary tagger
  - Based on handcrafted rules
- TnT tagger
  - Based on statistical modelling of sentences and their POS tags.
  - Hidden Markov Model tri-gram tagger
  - Trained on a large corpus of annotated sentences
Statistics: motivation

• Different tagging outcomes of the two taggers on the JOS corpus of 100k words

• Green: proportion of words where both taggers were correct

• Yellow: Both predicted the same, incorrect tag

• Blue: Both predicted incorrect but different tags

• Cyan: Amebis correct, TnT incorrect

• Purple: TnT correct, Amebis incorrect
<table>
<thead>
<tr>
<th>True</th>
<th>TnT</th>
<th>Amebis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preiskave</td>
<td>Ncfpn</td>
<td>Ncfpn</td>
</tr>
<tr>
<td>med</td>
<td>Si</td>
<td>Ncmsan</td>
</tr>
<tr>
<td>sodnim</td>
<td>Agumsi</td>
<td>Agumpd</td>
</tr>
<tr>
<td>postopkom</td>
<td>Ncmsi</td>
<td>Ncmpd</td>
</tr>
<tr>
<td>so</td>
<td>Va-r3p-n</td>
<td>Va-r3p-n</td>
</tr>
<tr>
<td>pokazale</td>
<td>Vmep-pf</td>
<td>Vmep-pf</td>
</tr>
</tbody>
</table>
Combining the taggers

Veža je smrdela po kuhanem zelju in starih, cunjastih predpražnikih.
Combining the taggers

A binary classifier; the two classes are TnT and Amebis

... prepričati italijanske pravosodne oblasti ...
Feature vector construction

Agreement features

\[
\begin{align*}
\text{POS}_{A-T} &= \text{yes}, \\
\text{Type}_{A-T} &= \text{yes}, \\
\ldots, \\
\text{Number}_{A-T} &= \text{yes}, \\
\text{Case}_{A-T} &= \text{no}, \\
\text{Animacy}_{A-T} &= \text{yes}, \\
\ldots, \\
\text{Owner\_Gender}_{A-T} &= \text{yes}
\end{align*}
\]

TnT features

\[
\begin{align*}
\text{POS}_T &= \text{Adjective}, \\
\text{Type}_T &= \text{general}, \\
\text{Gender}_T &= \text{feminine}, \\
\text{Number}_T &= \text{plural}, \\
\text{Case}_T &= \text{nominative}, \\
\text{Animacy}_T &= \text{n/a}, \\
\text{Aspect}_T &= \text{n/a}, \\
\text{Form}_T &= \text{n/a}, \\
\text{Person}_T &= \text{n/a}, \\
\text{Negative}_T &= \text{n/a}, \\
\text{Degree}_T &= \text{undefined}, \\
\text{Definiteness}_T &= \text{n/a}, \\
\text{Participle}_T &= \text{n/a}, \\
\text{Owner\_Number}_T &= \text{n/a}, \\
\text{Owner\_Gender}_T &= \text{n/a}
\end{align*}
\]

Amebis features

\[
\begin{align*}
\text{POS}_A &= \text{Adjective}, \\
\text{Type}_A &= \text{general}, \\
\text{Gender}_A &= \text{feminine}, \\
\text{Number}_A &= \text{plural}, \\
\text{Case}_A &= \text{accusative}, \\
\text{Animacy}_A &= \text{n/a}, \\
\text{Aspect}_A &= \text{n/a}, \\
\text{Form}_A &= \text{n/a}, \\
\text{Person}_A &= \text{n/a}, \\
\text{Negative}_A &= \text{n/a}, \\
\text{Degree}_A &= \text{undefined}, \\
\text{Definiteness}_A &= \text{n/a}, \\
\text{Participle}_A &= \text{n/a}, \\
\text{Owner\_Number}_A &= \text{n/a}, \\
\text{Owner\_Gender}_A &= \text{n/a}
\end{align*}
\]

This is the correct tag

⇒ label: Amebis

… prepričati italijanske pravosodne oblasti …
... italijanske pravosodne oblasti ...

(a) No context

... italijanske pravosodne oblasti ...

(b) Context
Classifiers

- **Naive Bayes**
  - Probabilistic classifier
  - Assumes strong independence of features
  - Black-box classifier

- **CN2 Rules**
  - If-then rule induction
  - Covering algorithm
  - Interpretable model as well as its decisions

- **C4.5 Decision Tree**
  - Based on information entropy
  - Splitting algorithm
  - Interpretable model as well as its decisions
Experiments: Dataset

- JOS corpus - approximately 250 texts (100k words, 120k if we include punctuation)
- Sampled from a larger corpus FidaPLUS
- TnT trained with 10 fold cross validation, each time training on 9 folds and tagging the remaining fold (for the meta-tagger experiments)
Experiments

• Baseline 1:
  – majority classifier (always predict what TnT predicts)
  – Accuracy: 53%

• Baseline 2
  – Naive Bayes
  – One feature only: Amebis full MSD
  – Accuracy: 71%
Baseline 2

- Naive Bayes classifier with one feature (Amebis full MSD) is simplified to counting the occurrences of two events for every MSD $f$:
  - #cases where Amebis predicted the tag $f$ and was correct: $n_c^f$
  - #cases where Amebis predicted the tag $f$ and was incorrect $n_w^f$
  - NB gives us the following rule, given a pair of predictions MSDa and MSDt: if $n_{c_{MSDa}} < n_{w_{MSDa}}$ predict MSDt, else predict MSDa.
Experiments: Different classifiers and feature sets

- Classifiers: NB, CN, C4.5
- Feature sets:
  - Full MSD
  - Decomposed MSD, agreement features
  - Basic features subset of the decomposed MSD features set (Category, Type, Number, Gender, Case)
  - Union of all features considered (full + decompositions)
- Scenarios:
  - no context
  - Context, ignore punctuation
  - Context, punctuation
Results

- Context helps
- Punctuation slightly improves classification
- C4.5 with basic features works best

### No context

<table>
<thead>
<tr>
<th>Feature set / Classifier</th>
<th>FULL TAG</th>
<th>DEC</th>
<th>BASIC</th>
<th>FULL+DEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>73.90</td>
<td>67.55</td>
<td>67.50</td>
<td>69.65</td>
</tr>
<tr>
<td>C4.5</td>
<td>73.51</td>
<td><strong>74.70</strong></td>
<td>74.23</td>
<td>73.59</td>
</tr>
<tr>
<td>CN2</td>
<td>60.61</td>
<td>72.57</td>
<td>71.68</td>
<td>70.90</td>
</tr>
</tbody>
</table>

### Context without punctuation

<table>
<thead>
<tr>
<th>Feature set / Classifier</th>
<th>FULL TAG</th>
<th>DEC</th>
<th>BASIC</th>
<th>FULL+DEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>73.10</td>
<td>68.29</td>
<td>67.96</td>
<td>70.55</td>
</tr>
<tr>
<td>C4.5</td>
<td>73.10</td>
<td><strong>79.23</strong></td>
<td>76.72</td>
<td></td>
</tr>
<tr>
<td>CN2</td>
<td>62.16</td>
<td>73.26</td>
<td>72.75</td>
<td>72.29</td>
</tr>
</tbody>
</table>

### Context with punctuation

<table>
<thead>
<tr>
<th>Feature set / Classifier</th>
<th>FULL TAG</th>
<th>DEC</th>
<th>BASIC</th>
<th>FULL+DEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>73.44</td>
<td>68.32</td>
<td>68.14</td>
<td>70.53</td>
</tr>
<tr>
<td>C4.5</td>
<td>74.18</td>
<td>78.91</td>
<td><strong>79.73</strong></td>
<td>77.68</td>
</tr>
<tr>
<td>CN2</td>
<td>62.23</td>
<td>74.27</td>
<td>72.82</td>
<td>73.01</td>
</tr>
</tbody>
</table>
Overall error rate

Amebis

TnT

Baseline 2 (Naive Bayes on only one feature; see Section 4.1)

Best (Context with punctuation, feature set BASIC)

Experimental setting

Overall error rate (%)
Thank you!