Automatic Enrichment of Croatian Morphological Lexicon Using Large Corpora and Web Search

Danijela Merkler, Željko Agić, Marko Tadić

Faculty of Humanities and Social Sciences
University of Zagreb

FASSBL-8, Dubrovnik, 2012-09-20
Motivation

- Manual enlargement of inflectional lexica is a time-consuming task requiring expertise
  - Assigning inflectional paradigms to potential entries
  - Ca 20 lemmas per hour
  - 10,000 lemmas equals 500 hours or ca 60 days of work

- Croatian Morphological Lexicon v 4.6
  - 110,000 lemmas, 4,000,000 entries (wordform, lemma, MSD)
  - Measured coverage: 96% on HNK and 91% on hrWaC
  - Lemmas added manually on daily basis
  - "Remaining" lemmas are expectedly infrequent

- Linguistically motivated rules for automatic enrichment
  - Derive female nouns (Ncf.*) from animate male nouns (Ncm.*y)
  - Derive possessive adjectives from male and female nouns (N.m.*y, subset of N.f.*)
  - Validation using hrWaC (and Google search index?)
Preliminary experiment

- used as a proof of concept
- manually selected 100 animate male nouns
- designed a set of derivation rules to produce female counterparts
  - 41 rule
  - 210 candidate lemmas generated
- queried hrWaC and Google for frequency evidence
- frequency ≥ 10

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male-female pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>hrWaC</td>
<td>95.24%</td>
<td>93.00%</td>
</tr>
<tr>
<td>Gold</td>
<td>92.38%</td>
<td>94.00%</td>
</tr>
<tr>
<td>Google</td>
<td>91.43%</td>
<td>86.00%</td>
</tr>
<tr>
<td></td>
<td>hrWaC</td>
<td>Google</td>
</tr>
</tbody>
</table>

Table: Preliminary experiment results
Extending experiment scope

- preliminary conclusions
  - lemmas identified with high accuracy
  - inflectional patterns assigned with derivation rules
  - hrWaC and Google scores comparable
  - Google Search API limited to 100 queries per day

- experiment extension
  - replace 100 manually selected nouns with real entries from the morphological lexicon
    - consider entries from other lexical sources
      - dictionaries, Croatian WordNet
  - include the possessive adjectives test case
  - use only hrWaC
  - data sparseness, frequency $\geq 2$
Experiment setup

- male-to-female test case
  - input nouns selected from the lexicon
    - filter Ncm.*y lemmas: 2.937 nouns
    - wordnet filtering not feasible (domain: person, SUMO: human, male)
    - dictionary does not denote animateness
  - 41 derivational rule
  - generated 6.810 candidate female nouns
    - 5.904 not covered by the lexicon
    - 1.713 confirmed by hrWaC, 985 not covered by the lexicon
  - evaluated both candidate lists (all not covered vs. confirmed not covered)
Experiment setup

- noun-to-adjective test case
  - input nouns selected from the lexicon
    - male filter: N.m.\*y
    - female filter: N.f.\* with specific inflectional patterns
    - included previously generated female nouns
    - input size: 12,950 candidate lemmas
  - 66 derivational rules
  - generated 6,583 candidate possessive adjectives
    - 6,486 not covered by the lexicon
    - 777 confirmed by hrWaC, 746 not covered by the lexicon
  - evaluated both candidate lists (all not covered vs. confirmed not covered)
Results

- substantial difference between accuracy on confirmed and unconfirmed female nouns
  - ambiguous suffixes in derivational rules
    - e.g. *ribič* → *ribička* (adjective lemma *ribički*)
  - high accuracy for both possessive adjective cases

<table>
<thead>
<tr>
<th>Test case</th>
<th>Count</th>
<th>Accuracy</th>
<th>New lemmas</th>
<th>New wordforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female conf’d</td>
<td>985</td>
<td>76%</td>
<td>750</td>
<td>10.500</td>
</tr>
<tr>
<td>Female all</td>
<td>5.904</td>
<td>27%</td>
<td>1.594</td>
<td>22.316</td>
</tr>
<tr>
<td>Adjective conf’d</td>
<td>746</td>
<td>98%</td>
<td>731</td>
<td>10.234</td>
</tr>
<tr>
<td>Adjective all</td>
<td>6.486</td>
<td>89%</td>
<td>5.773</td>
<td>80.822</td>
</tr>
<tr>
<td>Total</td>
<td>/</td>
<td>/</td>
<td>8.848</td>
<td>123.872</td>
</tr>
</tbody>
</table>

Table: Experiment results
Conclusions

- introduced 8,848 new lemmas to Croatian morphological lexicon
  - cleaning entries much faster than creating ones
  - saved ca 55 days of manual work
  - new version of the lexicon being prepared
    - includes these results and results of manual enlargement
    - expected ca 130,000 lemmas, more than 5,500,000 entries

- future work directions
  - guessing inflectional patterns for lemmas
  - including verb patterns
Thank you for your attention.