Retrieving Information in Croatian: building a simple and efficient rule-based stemmer

Nikola Ljubešić, research assistant
Damir Boras, full professor
Ozren Kubelka

INFuture2007

8 November 2007
problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)

two approaches in solving this problem

1. morphological lexicon - all lexemes with all possible word forms, very expensive to build
2. stemmer - a set of rules for a specific language (in English +ing, +ed, +s), far less expensive

in most complex systems both approaches - lexicon for known words and stemmer for unknown words

aim of paper twofold

1. analyze the practice in the IT industry in Croatia concerning the problem of Croatian morphology in information retrieval
2. construct a simple and efficient stemmer which could be easily used in the industry
Introduction

- problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)
- two approaches in solving this problem
Introduction

- problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)

- two approaches in solving this problem
  1. morphological lexicon - all lexemes with all possible word forms, very expensive to build

  2. stemmer - a set of rules for a specific language (in English +ing, +ed), far less expensive in most complex systems

  Both approaches - lexicon for known words and stemmer for unknown words

Aim of paper twofold

1. analyze the practice in the IT industry in Croatia concerning the problem of Croatian morphology in information retrieval

2. construct a simple and efficient stemmer which could be easily used in the industry
problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)

two approaches in solving this problem

1. morhological lexicon - all lexemes with all possible word forms, very expensive to build
2. stemmer - a set of rules for a specific language (in English +ing, +ed, +s), far less expensive
Introduction

- problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)

- two approaches in solving this problem
  1. morphological lexicon - all lexemes with all possible word forms, very expensive to build
  2. stemmer - a set of rules for a specific language (in English +ing, +ed, +s), far less expensive

- in most complex systems both approaches - lexicon for known words and stemmer for unknown words
Introduction

- problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)
- two approaches in solving this problem
  1. morphological lexicon - all lexemes with all possible word forms, very expensive to build
  2. stemmer - a set of rules for a specific language (in English +ing, +ed, +s), far less expensive
- in most complex systems both approaches - lexicon for known words and stemmer for unknown words
- aim of paper twofold
problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragogom, pretragama…)

two approaches in solving this problem

1. morphological lexicon - all lexemes with all possible word forms, very expensive to build
2. stemmer - a set of rules for a specific language (in English +ing, +ed, +s), far less expensive

in most complex systems both approaches - lexicon for known words and stemmer for unknown words

aim of paper twofold

1. analyze the practice in the IT industry in Croatia concerning the problem of Croatian morphology in information retrieval
Introduction

- problem of morphology while retrieving information - one lexeme has more word forms (pretraga, pretrage, pretrazi, pretragom, pretragama...)
- two approaches in solving this problem
  1. morphological lexicon - all lexemes with all possible word forms, very expensive to build
  2. stemmer - a set of rules for a specific language (in English +ing, +ed, +s), far less expensive
- in most complex systems both approaches - lexicon for known words and stemmer for unknown words
- aim of paper twofold
  1. analyze the practice in the IT industry in Croatia concerning the problem of Croatian morphology in information retrieval
  2. construct a simple and efficient stemmer which could be easily used in the industry
Tomislava Lauc’s PhD thesis - rules for noun and adjective paradigms, also for morphonological alternations, precision 90.26%, stemmer not evaluated on a corpus but on a lexicon.
Existing stemmers for Croatian

- Tomislava Lauc's PhD thesis - rules for noun and adjective paradigms, also for morphonological alternations, precision 90.26%, stemmer not evaluated on a corpus but on a lexicon.

 existing stemmers for Croatian

- Tomislava Lauc's PhD thesis - rules for noun and adjective paradigms, also for morphonological alternations, precision 90.26%, stemmer not evaluated on a corpus but on a lexicon
- Dobrica Pavlinušić - purpose of stemmer - help in retrieving texts of the official gazette of Republic of Croatia (Narodne novine), stemmer not evaluated, search engine at http://nn.rot13.org
- pogodak.hr uses probably a stemmer and a lexicon - not possible to confirm, business secret
Existing stemmers for Croatian

- Tomislava Lauc’s PhD thesis - rules for noun and adjective paradigms, also for morphonological alternations, precision 90.26%, stemmer not evaluated on a corpus but on a lexicon

- Dobrica Pavlinušić - purpose of stemmer - help in retrieving texts of the official gazette of Republic of Croatia (Narodne novine), stemmer not evaluated, search engine at http://nn.rot13.org

- pogodak.hr uses probably a stemmer and a lexicon - not possible to confirm, business secret

- Google started stemming Croatian documents in last few months, method unknown, most probably language independent - general rules + user feedback while using the search engine
acquiring insight in retrieval capabilities and technology used in top 30 Croatian web sites
Current situation in the industry

- acquiring insight in retrieval capabilities and technology used in top 30 Croatian web sites
- on-line questionnaire, out of 30 web administrators, 16 took the questionnaire
Current situation in the industry

- acquiring insight in retrieval capabilities and technology used in top 30 Croatian web sites
- on-line questionnaire, out of 30 web administrators, 16 took the questionnaire
- on question if the web site search engine is sensitive to morphological changes in Croatian - one positive answer
Current situation in the industry

- acquiring insight in retrieval capabilities and technology used in top 30 Croatian web sites
- on-line questionnaire, out of 30 web administrators, 16 took the questionnaire
- on question if the web site search engine is sensitive to morphological changes in Croatian - one positive answer
- most sites use just simple SQL queries (e.g., like) or search engines like Google
Current situation in the industry

- acquiring insight in retrieval capabilities and technology used in top 30 Croatian web sites
- on-line questionnaire, out of 30 web administrators, 16 took the questionnaire
- on question if the web site search engine is sensitive to morphological changes in Croatian - one positive answer
- most sites use just simple SQL queries (\texttt{\textasciitilde}, \texttt{like}) or search engines like Google
- 82\% believe a morphologically sensitive search engine is useful, 68\% of them are familiar with the method
Current situation in the industry

- acquiring insight in retrieval capabilities and technology used in top 30 Croatian web sites
- on-line questionnaire, out of 30 web administrators, 16 took the questionnaire
- on question if the web site search engine is sensitive to morphological changes in Croatian - one positive answer
- most sites use just simple SQL queries (=, like) or search engines like Google
- 82% believe a morphologically sensitive search engine is useful, 68% of them are familiar with the method
- 80% would upgrade their search engine with an open source of commercial stemmer, 10% would prefer an open source solution, 5% a commercial one, 5% are not interested
Building a simple and efficient stemmer

- up to this day, stemmers mostly too complex, built with no regard to specificities of IR, not properly evaluated
Building a simple and efficient stemmer

- up to this day, stemmers mostly too complex, built with no regard to specificities of IR, not properly evaluated
- optimization and evaluation datasets - portion of the Vjesnik on-line newspaper, validation and test set 10,000 randomly chosen articles (every cca. 4,5M tokens)
Building a simple and efficient stemmer

- up to this day, stemmers mostly too complex, built with no regard to specificities of IR, not properly evaluated
- optimization and evaluation datasets - portion of the Vjesnik on-line newspaper, validation and test set 10,000 randomly chosen articles (every cca. 4,5M tokens)
- validation set used to find most frequent nouns (query sample) and to extract rules

Ljubešić, Boras, Kubelka (INFuture2007)
Building a simple and efficient stemmer

- up to this day, stemmers mostly too complex, built with no regard to specificities of IR, not properly evaluated
- optimization and evaluation datasets - portion of the Vjesnik on-line newspaper, validation and test set 10,000 randomly chosen articles (every cca. 4,5M tokens)
- validation set used to find most frequent nouns (query sample) and to extract rules
- gold standard - built with help of a morphological lexicon (cca. 20K nouns) - for every possible query list of found tokens
Building a simple and efficient stemmer

- up to this day, stemmers mostly too complex, built with no regard to specificities of IR, not properly evaluated
- optimization and evaluation datasets - portion of the Vjesnik on-line newspaper, validation and test set 10,000 randomly chosen articles (every cca. 4,5M tokens)
- validation set used to find most frequent nouns (query sample) and to extract rules
- gold standard - built with help of a morphological lexicon (cca. 20K nouns) - for every possible query list of found tokens
- query sample - out of 13,261 nouns found in validation set, 1,033 manually checked, 33 removed because of the homonymy clash (bit, bilo, oko, toga, kada, dok etc.) - intuition about queries being mostly nouns and adjectives in basic form
Building a simple and efficient stemmer

- up to this day, stemmers mostly too complex, built with no regard to specificities of IR, not properly evaluated
- optimization and evaluation datasets - portion of the Vjesnik on-line newspaper, validation and test set 10,000 randomly chosen articles (every cca. 4.5M tokens)
- validation set used to find most frequent nouns (query sample) and to extract rules
- gold standard - built with help of a morphological lexicon (cca. 20K nouns) - for every possible query list of found tokens
- query sample - out of 13,261 nouns found in validation set, 1,033 manually checked, 33 removed because of the homonymy clash (bit, bilo, oko, toga, kada, dok etc.) - intuition about queries being mostly nouns and adjectives in basic form
- list of 32 suffix rules built by hand from the 1000 most frequent nouns
List of rules 1

(’, ’a’, ’u’, ’om’, ’i’, ’ima’, ’e’)
(’a’, ’e’, ’i’, ’u’, ’om’, ’ama’)
(’e’, ’a’, ’u’, ’em’, ’ima’)
(’o’, ’a’, ’u’, ’om’, ’ima’)

Ljubešić, Boras, Kubelka (INFuture2007)
List of rules 2

('t', 'ta', 'tu', 'tom', 'ti', 'ata', 'tima', 'te')
('zak', 'ska', 'sku', 'skom', 'sci', 'zaka', 'scima', 'ske')
('tak', 'tka', 'tku', 'tkom', 'tc', 'ci', 'taka', 'tcima', 'cima', 'tke')
('dac', 'ca', 'cu', 'cem', 'ci', 'daca', 'cima', 'ce')
('ga', 'ge', 'zi', 'gi', 'gu', 'gom', 'gama')
('st', 'sti', 'šću', 'stima')
('g', 'ga', 'gu', 'gom', 'zi', 'zima', 'ge')
('sao', 'sli', 'šlju', 'slima')
('t', 'ti', 'ću', 'tima')
('an', 'ni', 'nog', 'noga', 'nome', 'nomu', 'nim', 'ni', 'nih', 'nima', 'ne')
('o', 'og', 'oga', 'om', 'ome', 'omu', 'im', 'a', 'ih', 'ima', 'e')
('ni', 'an', 'nog', 'noga', 'nome', 'nomu', 'nim', 'ni', 'nih', 'nima', 'ne')
('a', 'e', 'oj', 'u', 'om', 'ih', 'ima')
('i', 'og', 'oga', 'om', 'ome', 'omu', 'im', 'ih', 'ima', 'e')
('i', 'og', 'oga', 'om', 'ome', 'omu', 'im', 'ih', 'ima', 'e')
Rule forms

- five different rule forms
Rule forms

- five different rule forms
  - there is a constraint regarding the last character of the invariable part of the lexeme (if consonant or vocal) (‘C’, ’a’, ’e’, ’i’, ’u’, ’om’, ’ama’) - zamka (zamke, zamki...) accepted, rastom (rasta, raste...) accepted, radia not
five different rule forms

1. there is a constraint regarding the last character of the invariable part of the lexeme (if consonant or vocal) (’C’, ’a’, ’e’, ’i’, ’u’, ’om’, ’ama’)
   - zamka (zamke, zamki...) accepted, rastom (rasta, raste...) accepted, radia not

2. there is a constraint regarding the first, entry suffix meaning that the query must fit the entry suffix for the rule to be applied to it (the entry suffix marked with square brackets) ([’a’], ’e’, ’i’, ’u’, ’om’, ’ama’), zamka accepted, radia accepted, rastom not
five different rule forms

1. there is a constraint regarding the last character of the invariable part of the lexeme (if consonant or vocal) (‘C’, ’a’, ’e’, ’i’, ’u’, ’om’, ’ama’) - *zamka* (*zamke, zamki*)... accepted, *rastom* (*rasta, raste*)... accepted, *radia* not

2. there is a constraint regarding the first, entry suffix meaning that the query must fit the entry suffix for the rule to be applied to it (the entry suffix marked with square brackets) ([’a’], ’e’, ’i’, ’u’, ’om’, ’ama’), *zamka* accepted, *radia* accepted, *rastom* not

3. there are both constraints from the first and the second form of rules (’C’, [’a’], ’e’, ’i’, ’u’, ’om’, ’ama’), *zamka* accepted, *rastom* and *radia* not

4. there are none of the constraints described above (’a’, ’e’, ’i’, ’u’, ’om’, ’ama’), *zamka*, *rastom* and *radia* accepted

5. there is no set of rules, but just a set of suffixes without any entry restrictions (”, ’a’, ’u’, ’om’, ’i’, ’ima’, ’e’, ’ama’, ’em’, ...), *zamka*, *rastom* and *radia* accepted
Rule forms

- five different rule forms
  1. there is a constraint regarding the last character of the invariable part of the lexeme (if consonant or vocal) (‘C’, ‘a’, ‘e’, ‘i’, ‘u’, ‘om’, ‘ama’) - zamka (zamke, zamki...) accepted, rastom (rasta, raste...) accepted, radia not
  2. there is a constraint regarding the first, entry suffix meaning that the query must fit the entry suffix for the rule to be applied to it (the entry suffix marked with square brackets) ([‘a’], ‘e’, ‘i’, ‘u’, ‘om’, ‘ama’), zamka accepted, radia accepted, rastom not
  3. there are both constraints from the first and the second form of rules (‘C’, [‘a’], ‘e’, ‘i’, ‘u’, ‘om’, ‘ama’), zamka accepted, rastom and radia not
  4. there are none of the constraints described above (‘a’, ‘e’, ‘i’, ‘u’, ‘om’, ‘ama’), zamka, rastom and radia accepted
five different rule forms

1. there is a constraint regarding the last character of the invariable part of the lexeme (if consonant or vocal) ('C', 'a', 'e', 'i', 'u', 'om', 'ama') - zamka (zamke, zamki...) accepted, rastom (rasta, raste...) accepted, radia not

2. there is a constraint regarding the first, entry suffix meaning that the query must fit the entry suffix for the rule to be applied to it (the entry suffix marked with square brackets) ([a], 'e', 'i', 'u', 'om', 'ama'), zamka accepted, radia accepted, rastom not

3. there are both constraints from the first and the second form of rules ('C', [a], 'e', 'i', 'u', 'om', 'ama'), zamka accepted, rastom and radia not

4. there are none of the constraints described above ('a', 'e', 'i', 'u', 'om', 'ama'), zamka, rastom and radia accepted

Optimization and evaluation

- using validation set - searching for set of suffixes which maximizes the $F_1$ measure regarding the gold standard - greedy optimization algorithm (steepest ascent hill climbing) - experimenting with different rule forms

<table>
<thead>
<tr>
<th>Rule Form</th>
<th>$F_1$ Measure</th>
<th>Precision</th>
<th>Recall</th>
<th># of Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9730</td>
<td>0.9775</td>
<td>0.9686</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>0.9747</td>
<td>0.9828</td>
<td>0.9667</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>0.9754</td>
<td>0.9774</td>
<td>0.9734</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>0.9417</td>
<td>0.9698</td>
<td>0.9151</td>
<td>19</td>
</tr>
</tbody>
</table>

Second form proven to be little bit better than others on test set, second form of rules achieves $F_1$ measure of 97.82%, precision 98.40%, recall 97.24%
Optimization and evaluation

- using validation set - searching for set of suffixes which maximizes the $F_1$ measure regarding the gold standard - greedy optimization algorithm (*steepest ascent hill climbing*) - experimenting with different rule forms

- $F_1$, precision and recall measures and number of rules on the validation set regarding the five forms of rules

<table>
<thead>
<tr>
<th>rule form</th>
<th>$F_1$</th>
<th>precision</th>
<th>recall</th>
<th># of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9730</td>
<td>0.9775</td>
<td>0.9686</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>0.9747</td>
<td>0.9828</td>
<td>0.9667</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>0.9754</td>
<td>0.9774</td>
<td>0.9734</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>0.9417</td>
<td>0.9698</td>
<td>0.9151</td>
<td>19</td>
</tr>
</tbody>
</table>

Second form proven to be little bit better than others on test set, second form of rules achieves $F_1$ measure of 97.82%, precision 98.40%, recall 97.24%.
Optimization and evaluation

- using validation set - searching for set of suffixes which maximizes the $F_1$ measure regarding the gold standard - greedy optimization algorithm (steepest ascent hill climbing) - experimenting with different rule forms

- $F_1$, precision and recall measures and number of rules on the validation set regarding the five forms of rules

<table>
<thead>
<tr>
<th>rule form</th>
<th>$F_1$</th>
<th>precision</th>
<th>recall</th>
<th># of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9730</td>
<td>0.9775</td>
<td>0.9686</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>0.9747</td>
<td>0.9828</td>
<td>0.9667</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>0.9754</td>
<td>0.9774</td>
<td>0.9734</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>0.9417</td>
<td>0.9698</td>
<td>0.9151</td>
<td>19</td>
</tr>
</tbody>
</table>

- second form proven to be little bit better than others
Optimization and evaluation

- using validation set - searching for set of suffixes which maximizes the $F_1$ measure regarding the gold standard - greedy optimization algorithm (*steepest ascent hill climbing*) - experimenting with different rule forms

$F_1$, precision and recall measures and number of rules on the validation set regarding the five forms of rules

<table>
<thead>
<tr>
<th>rule form</th>
<th>$F_1$</th>
<th>precision</th>
<th>recall</th>
<th># of rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9730</td>
<td>0.9775</td>
<td>0.9686</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td><strong>0.9781</strong></td>
<td>0.9843</td>
<td>0.9719</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>0.9747</td>
<td>0.9828</td>
<td>0.9667</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>0.9754</td>
<td>0.9774</td>
<td>0.9734</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>0.9417</td>
<td>0.9698</td>
<td>0.9151</td>
<td>19</td>
</tr>
</tbody>
</table>

- second form proven to be little bit better than others

- on test set, second form of rules acieves $F_1$ measure of 97,82%, precision 98,40%, recall 97,24%
$F_1$, precision and recall during *steepest ascent hill climbing* optimization algorithm with second rule form.
Adding rules for adjectives

- adjectives not main part of this phase of research
Adding rules for adjectives

- adjectives not main part of this phase of research
- adding six rules and observing the decline in precision and the growth of recall concerning the noun gold standard

<table>
<thead>
<tr>
<th># of rules</th>
<th>F1</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
</tr>
<tr>
<td>1</td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
</tr>
<tr>
<td>2</td>
<td>0.9779</td>
<td>0.9836</td>
<td>0.9722</td>
</tr>
<tr>
<td>3</td>
<td>0.9776</td>
<td>0.9831</td>
<td>0.9722</td>
</tr>
<tr>
<td>4</td>
<td>0.9773</td>
<td>0.9822</td>
<td>0.9725</td>
</tr>
<tr>
<td>5</td>
<td>0.9769</td>
<td>0.9814</td>
<td>0.9725</td>
</tr>
<tr>
<td>6</td>
<td>0.9763</td>
<td>0.9803</td>
<td>0.9725</td>
</tr>
</tbody>
</table>
Adding rules for adjectives

- adjectives not main part of this phase of research
- adding six rules and observing the decline in precision and the growth of recall concerning the noun gold standard

<table>
<thead>
<tr>
<th># of rules</th>
<th>F1</th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
</tr>
<tr>
<td>1</td>
<td>0.9781</td>
<td>0.9843</td>
<td>0.9719</td>
</tr>
<tr>
<td>2</td>
<td>0.9779</td>
<td>0.9836</td>
<td>0.9722</td>
</tr>
<tr>
<td>3</td>
<td>0.9776</td>
<td>0.9831</td>
<td>0.9722</td>
</tr>
<tr>
<td>4</td>
<td>0.9773</td>
<td>0.9822</td>
<td>0.9725</td>
</tr>
<tr>
<td>5</td>
<td>0.9769</td>
<td>0.9814</td>
<td>0.9725</td>
</tr>
<tr>
<td>6</td>
<td>0.9763</td>
<td>0.9803</td>
<td>0.9725</td>
</tr>
</tbody>
</table>

- decline in $F_1$ of just 0.18 - adjective and noun paradigms very simmilar
a XML-RPC web service written in Python, info how to use on http://faust.ffzg.hr/stemming/
Usage

- a XML-RPC web service written in Python, info how to use on http://faust.ffzg.hr/stemming/

Usage

```python
>>> import xmlrpclib
>>> server=xmlrpclib.Server('http://faust.ffzg.hr:8000')
>>> server.korijen('potraga')
['potragama', 'potragu', 'potrazi', 'potragi', 'potragom', 'potrage', 'potraga']
>>> server.korijen('pokret')
['pokretevima', 'pokrete', 'pokretovima', 'pokreteva', 'pokreta', 'pokretem', 'pokret', 'pokretevi', 'pokreti', 'pokretu', 'pokretovi', 'pokretima', 'pokretove', 'pokreću', 'pokretom', 'pokretova', 'pokreteve', 'pokreata']
```
a XML-RPC web service written in Python, info how to use on http://faust.ffzg.hr/stemming/

usage

```python
>>> import xmlrpclib
>>> server=xmlrpclib.Server('http://faust.ffzg.hr:8000')
>>> server.korijen('potraga')
['potragama', 'potragu', 'potrazi', 'potragi', 'potragom', 'potrage', 'potraga']
>>> server.korijen('pokret')
['pokretem', 'pokret', 'pokretevi', 'pokreti', 'pokretu', 'pokretovi', 'pokretima', 'pokretove', 'pokreću', 'pokretom', 'pokretova', 'pokreteve', 'pokreata']
```

plans to use the web service or the algorithm in Croatian academic directory http://www.hr
Usage

- a XML-RPC web service written in Python, info how to use on http://faust.ffzg.hr/stemming/

usage

```python
>>> import xmlrpclib
>>> server=xmlrpclib.Server('http://faust.ffzg.hr:8000')
>>> server.korijen('potraga')
['potragama', 'potragu', 'potrazi', 'potragi', 'potragom', 'potrage', 'potraga']
>>> server.korijen('pokret')
['pokretema', 'pokrete', 'pokretovima', 'pokreteva', 'pokreta', 'pokretem', 'pokret', 'pokretevi', 'pokreti', 'pokretu', 'pokretovi', 'pokretima', 'pokretove', 'pokreću', 'pokretem', 'pokretova', 'pokreteve', 'pokreata']
```

- plans to use the web service or the algorithm in Croatian academic directory http://www.hr

- an example web application for retrieving information from Vjesnik on-line using the web service
pronadeni oblici: potraga 82, potraga 44, potragu 28, potrage 27, potragama 1

U potrazi za djecom otetom iz Dječjeg doma pomogli i zagrebački taksisti
U potrazi za djecom otetom iz Dječjeg doma pomogli i zagrebački taksisti
» Josipovac « , u Nazorovoj 49 , nakon intenzivne potrage i niza operativnih mjera vraćeno je u srijedu u poslijepdnevnim M. V. ( 7 ) , zagrebačka policija je raspisala potragu . Nakon intenzivne potrage i opsežne pretrage terena , u
), zagrebačka policija je raspisala potragu . Nakon intenzivne potrage i opsežne pretrage terena , u srijedu oko 16.15 sati
), policija je odmah raspisala tjeralicu i započela intenzivnu potragu . Dva dana kasnije , u srijedu oko 10 sati
http://www.vjesnik.hr/html/2000/09/01/Clanak.asp?r=crn&c=4

» Policija me izbjegava , ako treba ići ću do Strasbourg «
izbjegava , ako treba ići ću do Strasbourg « Ni potraga niti očevid nisu obavljeni kako je trebalo , a od
neprestance tragali sinjskim krajem ne bi li pronašli djevojku . Potraga je završila 5. travnja kada je slučajni prolaznik Josip Bašić
osumnjičen . » izjavio je ogorčen Ivan Bešlić . U cilju potrage , a kasnije i istrage , Bešlić je sam javlja
urgirali u Ministarstvu započela je prava istraga , no niti potraga , niti očevid , nisu obavljeni kako je trebalo «

Suprug sumnja da mu je žena oteta i primorana na prostituciju
uredne dokumente , a za Jadrankom još nije bila raspisana potraga , nije bilo razloga za policijsku intervenciju . Upravo tu
Jedino što policiji zasad preostaje , je , nakon raspisane potrage po Hrvatskoj , raspisati i međunarodnu potragu .
, nakon raspisane potrage po Hrvatskoj , raspisati i međunarodnu potragu .
http://www.vjesnik.hr/html/2005/01/17/Clanak.asp?r=crn&c=3

Opozvana potraga za brodom sa 180 ilegalnih useljenika
Opozvana potraga za brodom sa 180 ilegalnih useljenika RIM , 13. veljače
13. veljače - Talijanska obalna straža u ponedjeljak je opozvala potragu za brodom , na kojem se smatra da je oko
Obalna straža iz talijanske luke Bari priopćila je da je potraga za nestalim brodom završila , te da nisu ništa pronašli
http://www.vjesnik.hr/html/2001/02/14/Clanak.asp?r=ssss&c=4
pronadeni oblici: pokret 173, pokret 121, pokretu 47, pokretima 16, pokreti 15, pokreću 15, pokrete 14, pokretom 7

Humanost bez granica

„što je i danas moto kojim se vodi Međunarodni pokret Crvenog križa i Crvenog polumjeseca. Stoga se rođendan tog radost življenja «. Mladi je Švicarac bio član filantropskog pokreta koji je pomagao starcima, bolesnicima i kažnjenicima. Pokret pokreta koji je pomagao starcima, bolesnicima i kažnjenicima. Pokret Budenje, komu je poslije pristupio, toliko je utjecao ta je povijesna heroina, čija je upornost i iznijedrila pokret medicinskih sestara, uspjela ranjenicima priskrbiti osnovnu medicinsku pomoć.

„što je i danas moto kojim se vodi Međunarodni pokret Crvenog križa i Crvenog polumjeseca. Stoga se rođendan tog radost življenja «. Mladi je Švicarac bio član filantropskog pokreta koji je pomagao starcima, bolesnicima i kažnjenicima. Pokret pokreta koji je pomagao starcima, bolesnicima i kažnjenicima. Pokret Budenje, komu je poslije pristupio, toliko je utjecao ta je povijesna heroina, čija je upornost i iznijedrila pokret medicinskih sestara, uspjela ranjenicima priskrbiti osnovnu medicinsku pomoć.

kako bi tome poučio građane. Načela Crvenog križa Načela pokreta ponajprije zahtijevaju sprječavanje i ublažavanje ljudske patnje, zaštite zdravlja
patnje, zaštite zdravlja i osiguranje poštivanja ljudskog dostojanstva. Pokret ne pravi razlike prema vjerskoj, nacionalnoj i klasnoj pripadnosti
službi nacionalnih vlada koja trebaju, u skladu s načelima pokreta, čuvati svoju autonomiju. Na diplomatskoj konferenciji koju je
http://www.vjesnik.hr/html/2006/05/11/Clanak.asp?r=pis&c=1

» Magičnim pokretima « korak bliže Castanedinim pričama

» Magičnim pokretima « korak bliže Castanedinim pričama Način na koji se Castaneda
širom svijeta održavaju tečajeve Tensegritya, sustava čarobnih i magičnih pokreta i konkretnog oblika praktične doktrine koji
raspřišlou energiju svakog čovjeka
ili slobodnih ratnika) koji jednom tijedno na Pantovčaku vježbaju pokrete. Mnogo je i onih, čuli smo, koji
. Mnogo je i onih, čuli smo, koji pokrete izvode individualno, a kroz radionicu je ukupno prošlo 200
. Radi se o praktičnom obliku drevnih znanja, seriji pokreta koji stimuliraju skupljanje raspršene energije i djeluju na mišiće i
), a prakticirali ih radi opće dobrobiti. Izvođenje pokreta priprema mišiće i tietve da brzo i učinkovito odgovaraju na
i tietve da brzo i učinkovito odgovaraju na duge serije pokreta koji zahtijevaju veću koncentraciju i kinestetičku memoriiju. - Ako
Conclusion

- stemming for IR in Croatian poorly used

- IT community recognizes the need for language technology in IR like building a simple stemmer out of 32 rules - greedy optimization algorithm with proper evaluation

- 97.82% F^1 concerning the gold standard (corpus-based, built with a lexicon)

- further research improve measure which penalizes over-generation

- real-world query sample - obtain a query sample from a search engine or directory (http://www.hr) and analyze it - use such sample in further research
Conclusion

- stemming for IR in Croatian poorly used
- IT community recognizes the need for language technology in IR like stemming
stemming for IR in Croatian poorly used

IT community recognizes the need for language technology in IR like stemming

building a simple stemmer out of 32 rules - greedy optimization algorithm with proper evaluation
stemming for IR in Croatian poorly used
IT community recognizes the need for language technology in IR like stemming
building a simple stemmer out of 32 rules - greedy optimization algorithm with proper evaluation
97.82% $F_1$ concerning the gold standard (corpus-based, built with a lexicon)
Conclusion

- stemming for IR in Croatian poorly used
- IT community recognizes the need for language technology in IR like stemming
- building a simple stemmer out of 32 rules - greedy optimization algorithm with proper evaluation
- 97.82% $F_1$ concerning the gold standard (corpus-based, built with a lexicon)
- further research
stemming for IR in Croatian poorly used

IT community recognizes the need for language technology in IR like stemming

building a simple stemmer out of 32 rules - greedy optimization algorithm with proper evaluation

97.82% $F_1$ concerning the gold standard (corpus-based, built with a lexicon)

further research
  - improve measure which penalizes over-generation
Conclusion

- stemming for IR in Croatian poorly used
- IT community recognizes the need for language technology in IR like stemming
- building a simple stemmer out of 32 rules - greedy optimization algorithm with proper evaluation
- 97.82% $F_1$ concerning the gold standard (corpus-based, built with a lexicon)
- further research
  - improve measure which penalizes over-generation
  - real-world query sample - obtain a query sample from a search engine or directory (http://www.hr) and analyze it - use such sample in further research