langnet

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
Tel.: + 385 51 584 700 Fax: + 385 51 584 749
www.langnet.uniri.hr

# Toward Network-based Keyword Extraction from Multitopic Web Documents

**Ana Meštrović**
*amestrovic@uniri.hr*
**Sanda Martinčić-Ipšić**
*smarti@uniri.hr*
**Sabina Šišović**
*ssisovic@uniri.hr*

@


**ITIS 2014 – November 5-7 in Šmarješke toplice, Slovenia**

# Introduction

langnet

- **Keyword extraction**
  - problem of **automatic identification** of the **important terms** or **phrases** in text documents
    - most salient features in text
  - numerous applications: information retrieval, automatic indexing, text summarization, semantic description and classification
  - **unsupervised** graph-based keyword extraction algorithms (unsupervised approach, based on centrality measures)
  - Initial experiment (the data source – not annotated)

# Centrality measures

langnet

## degree, closeness and betweenness centrality

$$dc_i^{(in/out)} = \frac{k_i^{(in/out)}}{N-1} \qquad cc_i = \frac{N-1}{\sum_{i \neq j} d_{ij}} \qquad bc_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}$$

## strength and selectivity

$$s_i^{in/out} = \sum_j w_{ji/ij} \qquad e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}$$

## selectivity

- average node weight

# Datasets

- 4 different web sources:
  - business portal Gospodarski list (GL),
  - legislative portal Narodne novine (NN),
  - news portal with forum Index.hr (IN),
  - daily newspaper portal Slobodna Dalmacija (SD)

| Dataset | GL | NN | IN | SD |
|---|---|---|---|---|
| Number of words | 199417 | 146731 | 118548 | 44367 |
| Number of nodes, *N* | 22727 | 13036 | 15065 | 9553 |
| Number of links, *K* | 105171 | 55661 | 28972 | 25155 |

# Network construction

langnet

- directed and weighted co-occurrence networks
  - **node**: word from the text
  - **link**: words that are linked if they are direct neighbours in a sentence
  - **weight on the link**: number of co-occurrence of the pair of words in the text that are connected with that link

- **Python + NetworkX**

# Experiment

langnet

- The first part of the experiment:
  - computing: in/out-degree, closeness, betweenness centrality and selectivity for each node
  - **ranking** all nodes (words) according to the values of each of these measures, obtaining **top 10 keyword candidates**

- The second part of the experiment:
  - computing **in-selectivity** and **out-selectivity** for node
  - **ranking** nodes according to the highest in/out-selectivity values
  - detecting **neighbor nodes** with the highest (in/out) weight for each node -> **word tuples**

# Experiment II

langnet

- The third part of the experiment:
  - applying **different filters** on the in/out-selectivity based word tuples
    - the **stopwords-filter**: we filter out all tuples that contain stopwords
    - the **high-weights-filter**: from the in/out-selectivity based word tuples we select only those tuples that have the same values for the selectivity and weight
    - the **combination** of the first two filters

# Results I

langnet

- Top ten words ranked according to the diferent measures (NN dataset)

| selectivity | in-degree | out-degree | closeness | betwenness |
|---|---|---|---|---|
| novine | i | i | i | i |
| temelju | u | u | ili | u |
| manjinu | za | je | je | za |
| srpsku | na | za | se | ili |
| skladu | ili | se | da | na |
| snagu | iz | ili | usluga | je |
| osiguranju | te | na | zakona | se |
| narodnim | je | o | a | o |
| novinama | se | te | skrbi | te |
| kriza | s | članak | hrta | iz |

# Results II

- Top ten ranked in/out-selectivity based word-tuples

| word before | word | e-in | w | word | word after | e-out | w |
|---|---|---|---|---|---|---|---|
| narodne | novine | 326 | 326 | srpsku | nacionalnu | 222 | 222 |
| na | temelju | 317 | 317 | nacionalnu | pripadnost | 183 | 1 |
| nacionalnu | manjinu | 275 | 2 | ovjesne | jedrilice | 159 | 159 |
| za | srpsku | 222 | 222 | narodnim | novinama | 129 | 129 |
| u | skladu | 202 | 202 | narodne | jazz | 111 | 1 |
| na | snagu | 172 | 172 | manjinu | gradu | 78 | 1 |
| o | osiguranju | 134 | 43 | ovoga | sporazuma | 72 | 1 |
| u | narodnim | 129 | 129 | crvenog | kristala | 72 | 3 |
| narodnim | novinama | 129 | 129 | skladu | provjeriti | 67 | 1 |
| crvenoga | križa | 99 | 2 | oružanih | sukoba | 58 | 4 |

# Results III

langnet

- Top ten ranked word-tuples **without stopwords**

| word before | word | e-in | w | word | word after | e-out | w |
|---|---|---|---|---|---|---|---|
| narodne | novine | 326 | 326 | srpsku | nacionalnu | 222 | 222 |
| nacionalnu | manjinu | 275 | 2 | nacionalnu | pripadnost | 183 | 1 |
| narodnim | novinama | 129 | 129 | ovjesne | jedrilice | 159 | 159 |
| crvenoga | križa | 99 | 2 | narodnim | novinama | 129 | 129 |
| jedinicama | regionalne | 65 | 1 | narodne | jazz | 111 | 1 |
| nacionalne | manjine | 61 | 61 | manjinu | gradu | 78 | 1 |
| rizika | snaga | 57 | 1 | ovoga | sporazuma | 72 | 1 |
| medije | ubroj | 47 | 1 | crvenog | kristala | 72 | 3 |
| crveni | križ | 42 | 42 | skladu | provjeriti | 67 | 1 |
| uopravni | spor | 41 | 41 | oružanih | sukoba | 58 | 4 |

# Results IV

langnet

- Top ten word-tuples **with equal in/out selectivity**

| word before | word | e-in=e-out | word | word after | e-in=e-out |
|---|---|---|---|---|---|
| narodne | novine | 326 | srpsku | nacionalnu | 222 |
| na | temelju | 317 | ovjesne | jedrilice | 159 |
| za | srpsku | 222 | narodnim | novinama | 129 |
| u | skladu | 202 | sjedištem | u | 55 |
| na | snagu | 172 | objavit | će | 53 |
| u | narodnim | 129 | republici | hrvatskoj | 52 |
| narodnim | novinama | 129 | albansku | nacionalnu | 52 |
| i | dopunama | 68 | republika | hrvatska | 49 |
| nacionalne | manjine | 61 | oplemenjivačkog | prava | 45 |
| sa | sjedištem | 55 | mađarsku | nacionalnu | 40 |

# Results V

- Top ten word-tuples **with equal in/out selectivity and without stopwords** (combination)

| word before | word | e-in=e-out | word | word after | e-in=e-out |
|---|---|---|---|---|---|
| narodne | novine | 326 | srpsku | nacionalnu | 222 |
| narodnim | novinama | 129 | ovjesne | jedrilice | 159 |
| nacionalne | manjine | 61 | narodnim | novinama | 129 |
| crveni | križ | 42 | republici | hrvatskoj | 52 |
| upravni | spor | 41 | albansku | nacionalnu | 52 |
| ovjesnom | jedrilicom | 39 | republika | hrvatska | 49 |
| elektroničke | medije | 36 | oplemenjivačkog | prava | 45 |
| nacionalnih | manjina | 35 | mađarsku | nacionalnu | 40 |
| domovinskog | rata | 33 | romsku | nacionalnu | 33 |
| ivan | vrljić | 30 | nadzorni | odbor | 33 |

# Conclusion

langnet

- Selectivity-based keyword extraction
  - better results than other three centrality measures
  - extraction of word tuples: phrases, names
  - robust to noise
  - capable of extrating from multitopic datasurces

- Future work:
  - evaluation of the results
  - experiment with lammatised texts
  - experiment with different network measures

# Toward Network-based Keyword Extraction from Multitopic Web Documents

langnet

?

**Ana Meštrović**

*amestrovic@uniri.hr*

Department of Informatics, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia