

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

Network-Enabled Keyword Extraction for Under-Resourced Languages

Slobodan Beliga
sbeliga@inf.uniri.hr
Sanda Martinčić-Ipšić
smarti@inf.uniri.hr





IKC 2016 – Cluj-Napoca
8th September 2016

Introduction 1/2

langnet

Keyword extraction (KE)

automatically identify a **set of terms** that best describe the document

identify and rank the most representative features of the source

applications in text:

- summarization, indexing, labeling, categorization, clustering

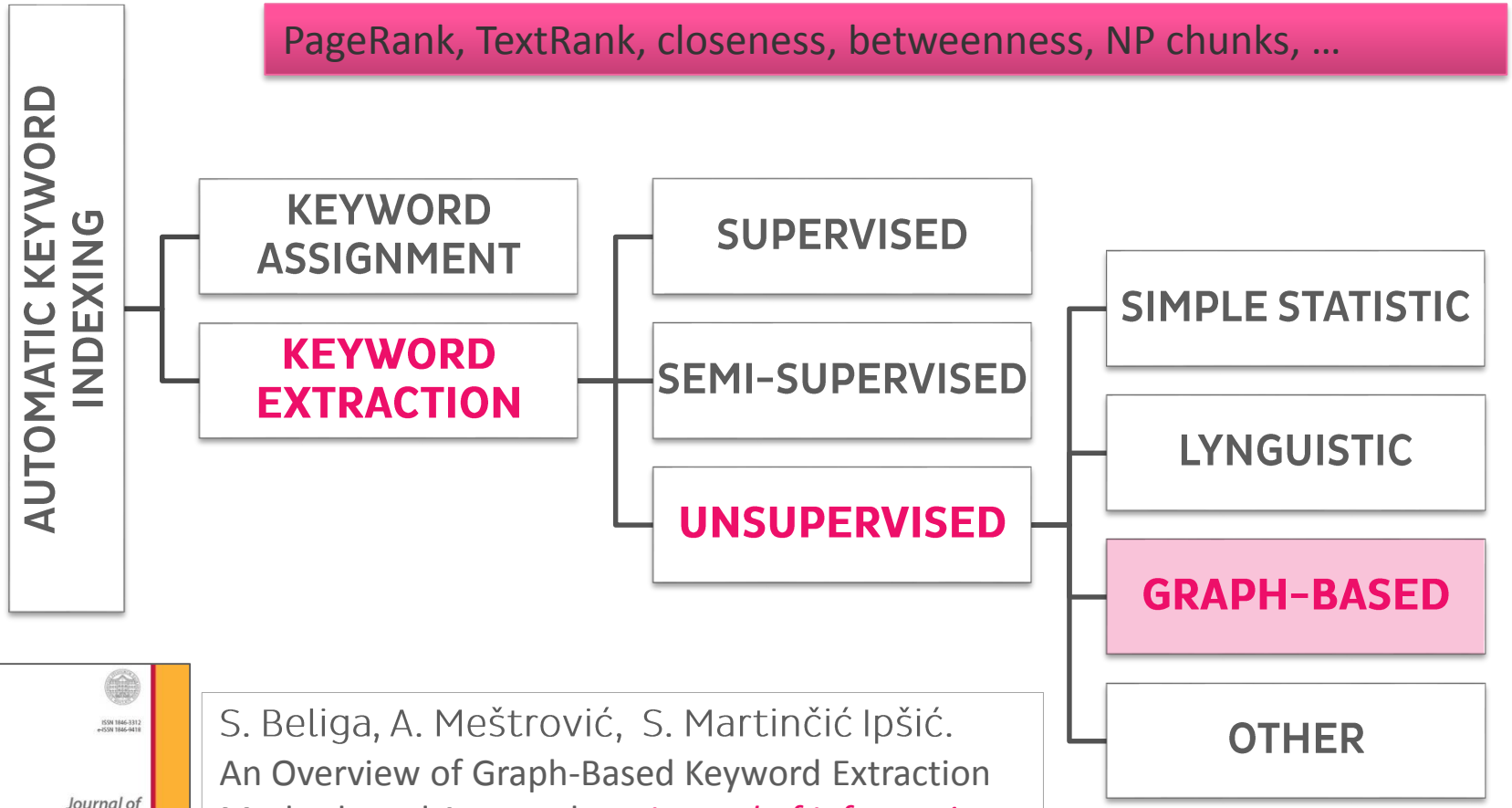
KE methods can be roughly divided into 3 categories:

- supervised,
- semi-supervised and
- unsupervised

Introduction 1/2

langnet

PageRank, TextRank, closeness, betweenness, NP chunks, …





S. Beliga, A. Meštrović, S. Martinčić Ipšić. An Overview of Graph-Based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, vol. 39, no. 1, 2015.

Introduction 2/2

langnet

automatic KE still remains open question

- especially for content written in **under-resourced languages**
  - there are no reliable tools which can be used for KE task and text preprocessing
  - POS, MSD taggers, stemmers, lemmatisers, stop-word lists
  - lexical resources: Wordnet, controlled vocabularies, benchmark or monitoring datasets

aim of this work

- discuss the problems of keyword extraction in under-resourced languages
- recommend possible solution: network or **graph-enabled KE methods**
  - these methods use knowledge incorporated in the structure of network or graph to extract keywords
  - way to circumvent unavailable linguistic tools required in a certain KE method development
- provide a list of available benchmark datasets for KE development and evaluation in order to illustrate the problem of the lack of resources

Under-Resourced Languages 1/5



There are more than **7000 languages** today in the world (Ethnologue, 2016)

- small fraction of them is supported with the resources required for implementation of NLP technologies (Besacier et al., 2014)

Main stream NLP is mostly concerned with languages

- for which large resources are available or
- which have suddenly become of concern because of the **economic interest** or **political influence**

The term "**under-resourced languages**" was introduced by **Krauwer** (2003) and complemented by **Berment** (2004)

- Different terminology:
  - low-density;
  - less-resourced;
  - under-resourced

Under-Resourced Languages 2/5

langnet

Under-resourcenest criteria

1. **Krauwer** (2003) and **Berment** (2004)
   - define criteria to consider a particular language as u-r:
     - lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, etc.
2. **BLARK** (Basic Language Resource Kit) **concept**
   - defined as the minimal set of language resources that is necessary to do any precompetitive research and education at all
   - it must be under 10/20 in order to confirm affiliation to u-r languages
   - comprises criteria, such as:
     - written language corpora, mono and bilingual dictionaries, terminology collections, grammars, annotation standards and tools, corpus exploration and exploitation tools, different modules (e.g. taggers, morphological analyzers, parsers, speech recognizers, TTS)

Under-Resourced Languages 3/5

langnet

Berment define Human languages categorization:

- based on their digital "readiness" or presence in cyberspace and software tools:
  - "tau"-languages: totally-resourced languages,
  - "mu"-languages: medium-resourced languages, and
  - "pi"-languages: under-resourced languages.

→ individual researchers

META classifies language technology into the following areas:

- Machine Translation,
- Speech Processing,
- Text Analysis, and
- Speech and Text Resources.

→ group of researchers, and commercial technology providers, private and corporate language technology users, language professionals

Within these areas languages are classified into the following categories: **excellent, good, moderate, fragmentary and weak/no support.**

Under-Resourced Languages 4/5

langnet

Text Analysis is the most important area for KE

- languages marked with '**+**' have the lowest support in **weak support category**
- presented data are related exclusively to European languages

Language	Machine Translation	Speech Processing	Text Analysis	Speech and Text Resources
Bulgarian	+			
Croatian	+	+	+	
Czech	+			
Danish	+			
Estonian	+		+	
Finish	+			
Greek	+			
Icelandic	+	+	+	+
Irish	+		+	+
Latvian	+	+	+	+
Lithuanian	+	+	+	+
Maltese	+	+	+	+
Portuguese	+			
Serbian	+		+	
Slovak	+			
Slovene	+			
Swedish	+			
Welsh	+	+	+	+

Under-Resourced Languages 5/5

- Besides to **weak support** category, there are other languages that are classified into **fragmentary** and few of them in **moderate** category.
- **English** is the only language with good support in all areas
- Expressed in the proportions:
  - weak supported - 30%,
  - fragmentary supported - 16.66%,
  - moderate supported: 16.66% and
  - good supported - 3.33%.

Language	Good	Moderate	Fragmentary
English	+		
Dutch		+	
French		+	
German		+	
Italian		+	
Spanish		+	
Basque			+
Bulgarian			+
Catalan			+
Czech			+
Danish			+
Finnish			+
Galician			+
Greek			+
Hungarian			+
Norwegian			+
Polish			+
Portuguese			+
Romanian			+
Slovak			+
Slovene			+

Problems in KE and Motivation 1/2



There are numerous KE methods for richer-resourced languages with remarkable performance

- these methods are most often developed for the English language
- in the absence of language tools it is difficult to adopt them for other languages, especially for u-r languages
- limited language scalability (portability)

In order to support **multilingualism** we propose unsupervised methods, **network-enabled methods** for KE.

- Network structure enables representation of the input text as graph or network, regardless of language.

Problems in KE and Motivation 2/2

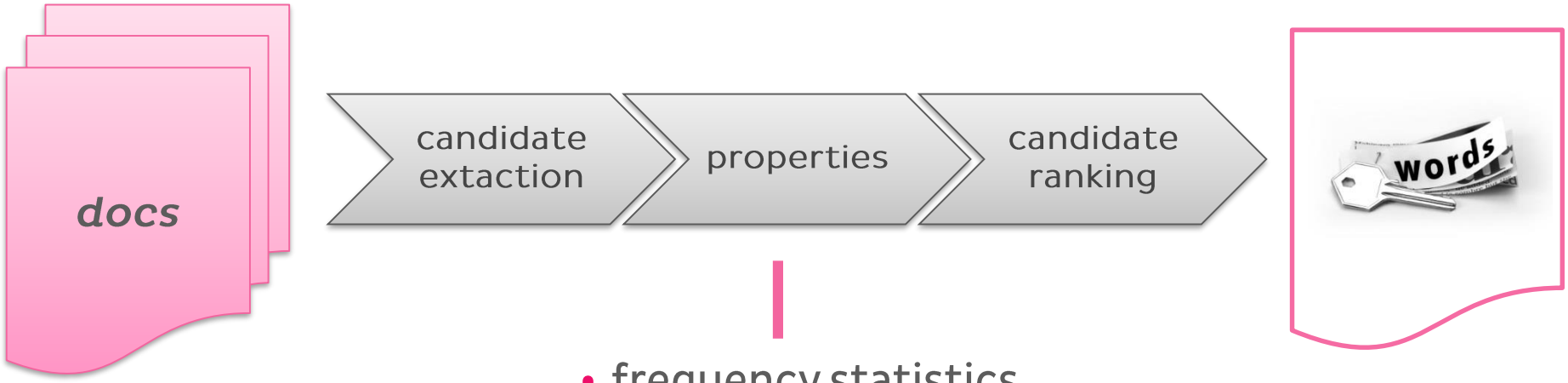
langnet

Our solution

- instead of developing new tools for language of interest, use graph-based method
  - Selectivity-Based Keyword Extraction (**SBKE**)

- application of this method requires only tuning of various parameters which are inherent for particular language:
  - fine tuning of parameters for candidate extraction,
  - setting the filtering thresholds for keyword expansion, ...

How KE works?



- frequency statistics
- occurence positions
- keyphraseness
- semantic relatedness
- other features

Network-Enabled KE Concept

langnet

network-based approach

- or graph, since the number of words in isolated documents is limited

- the source (document, text) is modelled in a network
  - words are nodes and their co-occurrence is represented with links
- reason to use co-occurrence, and not any semantic or syntactic relation is the **lack of language tools** which could extract these relations
- enables the exploration of the relationships and structural information incorporated in text very efficiently
- the **keyword extraction** can exploit: **network, subnetwork** or **node** level measures:
  - **selectivity –** the average strength of the node:

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}$$

Co-occurrence network construction

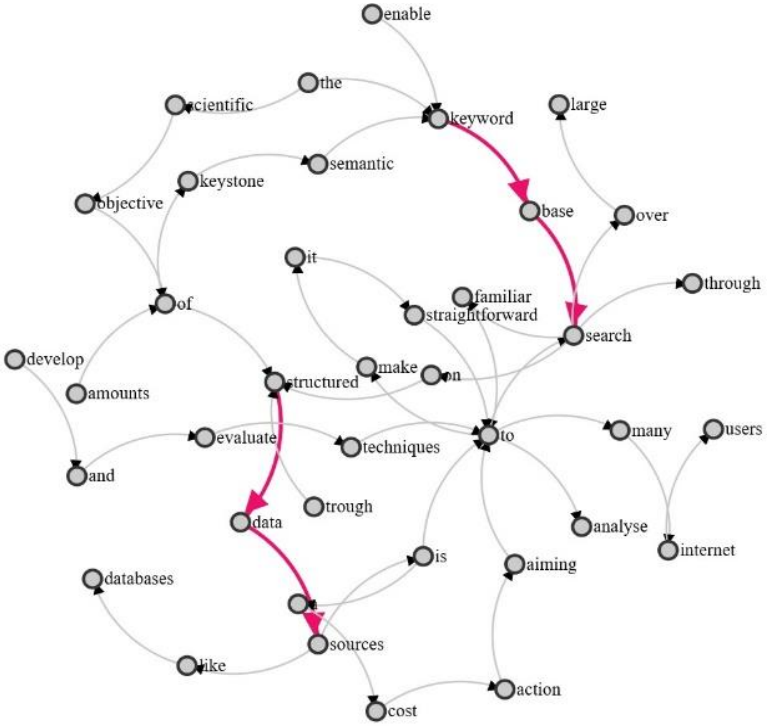
langnet

TEXT: *“KEYSTONE - semantic keyword-based search on structured data sources" is a COST Action aiming to make it straightforward to search through structured data sources like databases using the keyword-based search familiar to many internet users. The scientific objective of KEYSTONE is to analyse, design, develop and evaluate techniques to enable keyword-based search over large amounts of structured data.*

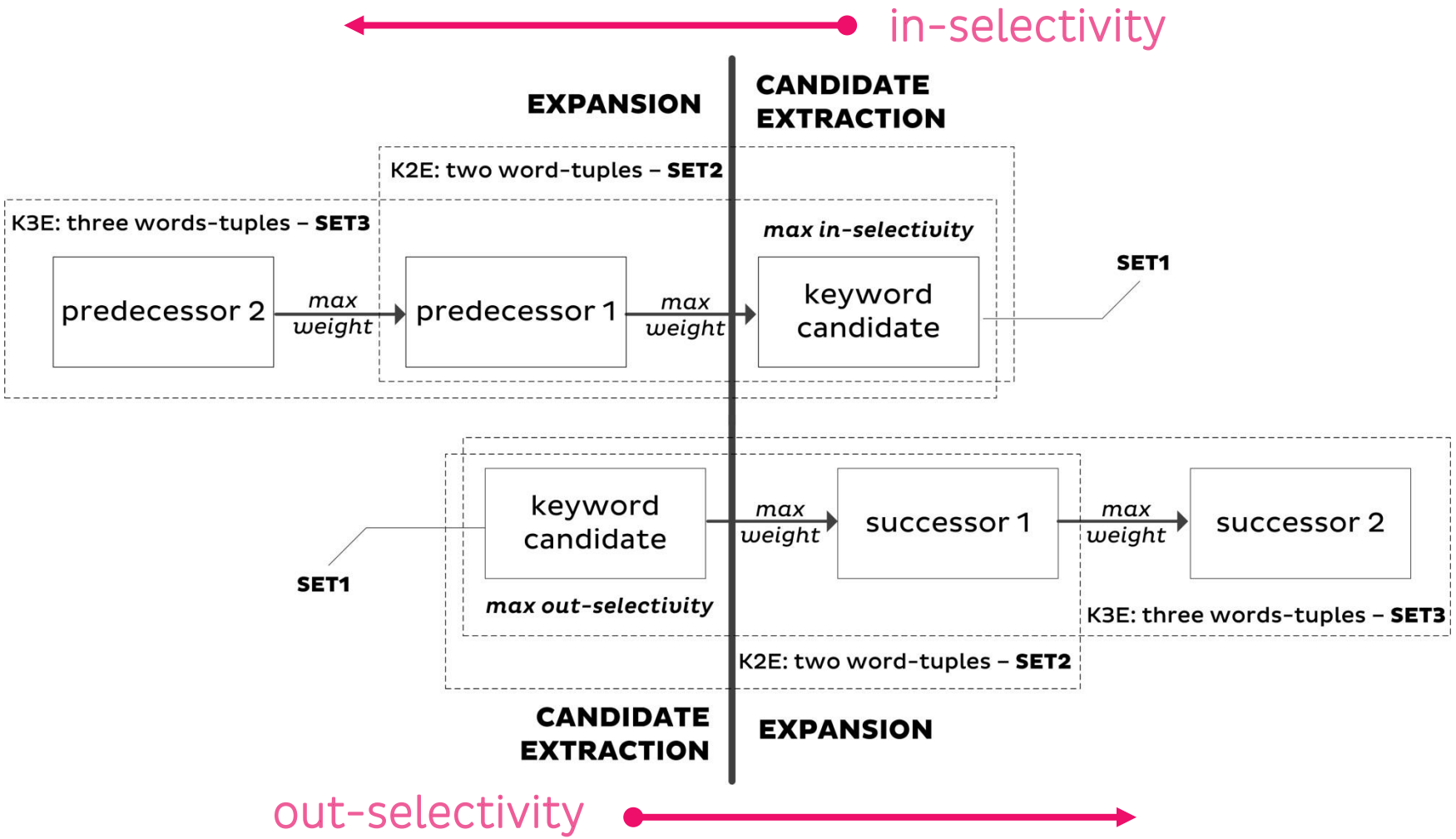
directed and weighted co-occurrence networks:

node: each word
links: two words are linked if they are adjacent in the sentence
weights: proportional to the overall co-occurrence frequencies of the corresponding word pairs

implemetation: Python + NetworkX



SBKE method


in-selectivity
EXPANSION
CANDIDATE EXTRACTION
K2E: two word-tuples – SET2
K3E: three words-tuples – SET3
max in-selectivity
predecessor 2
max weight
predecessor 1
max weight
keyword candidate
SET1
SET1
keyword candidate
max weight
successor 1
max weight
successor 2
max out-selectivity
K3E: three words-tuples – SET3
K2E: two word-tuples – SET2
CANDIDATE EXTRACTION
EXPANSION
out-selectivity

SBKE Method Portability for Under-Resourced Languages

langnet



SBKE method

- does not require linguistic knowledge
- **purely** from **statistical and structural information**
- main advantages:
  - networks are constructed from pure co-occurrence of words in the input texts
  - suitable for many European under-resourced languages
  - the method achieves results which are **above the TFxIDF baseline** for **English** and under-resourced **Croatian** language
  - **Preprocessing is not necessary** because SBKE works without stemming or lemmatization

S. Beliga, A. Meštrović, S. Martinčić-Ipšić. "Selectivity-Based Keyword Extraction Method". *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol.12, no.3, pages 1-26, 2016.

Textual Resources for KE in Under-Resourced Languages 1/2



benchmark datasets which contain keywords annotated by human experts

- most of the available datasets are in the
  - **English** language (12 datasets)
- other available datasets cover the
  - **French** (DEFT – scientific articles published in social science journals);
  - **French** and **Spanish** (FAO 780 – FAO publications with Agrovoc terms);
  - **Polish** (abstracts of academic papers for PKE method),
  - **Portuguese** (tweet dataset and news transcriptions) (2 datasets)
  - **Croatian** (HINA – news articles).
- other u-r languages which are in our focus do not have developed datasets for keyword extraction task

Textual Resources for KE in Under-Resourced Languages 2/2

benchmark datasets – characteristics

- datasets with controlled vocabulary
  - consist of **manually annotated keywords** by humans using only words from original text,
    - titles of Wikipedia articles or
    - some predefined list of allowed words which can be used as keywords
  - **important for different supervised / unsupervised approaches**
    - such datasets are particularly suitable for methods which are not able to generate new words
- number of human annotators
    - subjective task and gold standard
    - the quality of the dataset is higher if the number of human (individuals or teams) annotators is higher

Conclusion

langnet

This work briefly

- describes graph or network-enabled keyword extraction methods
- explains why these methods are suitable for under-resourced languages
- provides the detailed list of datasets for keyword extraction for European languages
- shows that there is not enough datasets for KE task
- emphasizes the importance of KE develompment for under-resourced languages
  - open the possibilities for the development of other applications which in its initial phase require keywords (automatic summarization, high level semantic description, indexing, ...)

SBKE – Serbian language

www.novosti.rs

- preliminary results
- short news articles on Serbian language

	TITLE	KEYWORDS (translated to English)
POLITICS	Migrants	refugees, political life, Angela Merkel, elections, united, more, Austria, options, Germany, population, year
ECONOMICS	Credit without a permanent job	customers, credit, capable, banks, ability, loan, institutions, interest rates, reserve, categories, evaluation, mandatory contract, criteria, agent
SPORTS	Serbian paralympic athletes traveled to the Rio	athlete, Rio, support, champion, medal, pride, Minister, preparation, effort, table tennis, team, London, Uroš Zeković

Future work

Try SBKE method for other under-resourced languages to show that

- knowledge incorporated in the network should replace non-existing linguistic tools necessary for keyword extraction from semi-structured web sources

- we will focus on:
  - KE dataset development for u-r languages
  - Croatian, Serbian, Estonian, Latvian, Lithuanian and Maltese

Network-Enabled Keyword Extraction for Under-Resourced Languages





langnet

?

Slobodan Beliga
sbeliga@inf.uniri.hr
Sanda Martinčić-Ipšić
smarti@uniri.hr

Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
www.langnet.uniri.hr

References

1. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: An Overview of Graph-Based Keyword Extraction Methods and Approaches, Journal of Information and Organizational Sciences, 39(1), 1-20 (2015)
2.Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic Speech Recognition for Under-resourced Languages: A Survey. Speech Communication, vol. 56, 85-100, (2014)
3.Krauwer, S.: The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003, pp. 8-15, Moscow, Russia (2003)
4.Berment, V.: Méthodes pour informatiser des langues et des groupes de langues "peu dotées". PhD Thesis, J. Fourier University – Grenoble I, (2004)
5.Abilhoa, W. D., & Castro, L. N.: A keyword extraction method from twitter messages represented as graphs. Applied Mathematics and Computation, 240, 308-325 (2014)
6.Palshikar, G. K.: Keyword Extraction from a Single Document Using Centrality Measures. Pattern Recognition and Machine Intelligence (LNCS) Second International Conference, PReMI 2007, Springer Berlin Heidelberg, vol. 4815, pp. 503-510 (2007)
7.Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In Proceedings of Empirical Methods in Natural Language Processing – EMNLP 2004. Barcelona, Spain: ACL, pp. 404-411 (2004)
8.META-NET – official site (May 2016), http://www.meta-net.eu/
9.META-NET White Paper Series: Key Results and Cross-Language Comparison (May 2016), http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison
10.Joorabchi, A., Mahdi, A. E.: Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. Journal of Information Science, 39(3), 410–426 (2013)
11.Lahiri, S., Choudhury, S.R., Caragea, C.: Keyword and Keyphrase Extraction Using Cen-trality Measures on Collocation Networks. arXiv preprint arXiv:1401.6571, (2014)
12.Grineva, M., Grinev, M., Lizorkin, D.: Extracting key terms from noisy and multitheme documents. ACM 18th conference on World Wide Web, pp.661–670, (2009)
13.Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Toward Selectivity-Based Keyword Extraction for Croatian News". CEUR Proceedings of the Workshop on Surfacing the Deep and the Social Web (SDSW 2014), Vol. 1310, pp. 1-8, Riva del Garda, Trentino, Italy (2014)
14.Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Selectivity-Based Keyword Extraction Meth-od. International Journal on Semantic Web and Information Systems (IJSWIS), 12(3) (2016) (accepted)
15.Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, ACL-IJCNLP 2015, Beijing, China (2015)
16.Paroubek, P., Zweigenbaum, P., Forest, D., Grouin, C.: Indexation libreet controlee d'articles scientifiques. Presentation et resultats du defi fouille de textes DEFT2012. In Proceedings of the DEfi Fouille de Textes 2012 Workshop, pp. 1–13, (2012)
17.Kozłowski, M.: PKE: a novel Polish keywords extraction method. Pomiary Automatyka Kontrola, R.60(5), 305-308 (2014)
18.Mijić, J., Dalbelo-Bašić, B., Šnajder, J.: Robust Keyphrase Extraction for a Large-Scale Croatian News Production System. In Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Balkan Languages, Zagreb, Croatia: Croatian Language Technologies Society, pp. 59-66 (2010)
19.Collection of comparable Lithuanian, Latvian and Estonian laws and legislations (June 2016), http://metashare.nlp.ipipan.waw.pl/metashare/repository/browse/collection-of-comparable-lithuanian-latvian-and-estonian-laws-and-legisla-tions/8d0d633eae7711e2a28e525400c0c5ef33b6cfc6ca074e1ab58859157c8374e7/#
20.Zunde, P., Dexter M. E.: Indexing consistency and quality. American Documentation, 20(3), 259–267 (1969)
21.Loza, V., Lahiri, S., Mihalcea, R., Lai, P.: Building a Dataset for Summarization and Keyword Extraction from Emails. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 2441-2446, Reykjavik, Iceland, (2014)
22.Su, N. K., Medelyan, O., Min-Yen, K., Timothy, B.: Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation. 47(3), 723-742 (2013)
23.Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., et al.: Part-of-speech tagging for twitter: annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, HLT '11, Stroudsburg, PA, USA. Association for Computational Linguistics (2011)
24.Marujo, L., Wang, L., Trancoso, I., Dyer, C., Black, A. W., Gershman, A., et al.: Automatic Keyword Extraction on Twitter. ACL (2015).
25.Medelyan, O. Human-competitive automatic topic indexing. PhD thesis. Department of Computer Science, University of Waikato, New Zealand (2009)
26.Hulth, A.: Improved Automatic Keyword Extraction Given More Linguistic Knowledge. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 216-223 (2003)
27.Nguyen, T. D., Kan, M.: Key phrase Extraction in Scientific Publications. Proceeding of International Conference on Asian Digital Libraries, pp. 317-326 (2007)
28.Wan, X., Xiao, J.: CollabRank: towards a collaborative approach to single-document keyphrase extraction. In Proceedings of COLING, pp. 969–976 (2008)
29.Krapivin, M., Autaeu, A., Marchese, M.: Large dataset for keyphrase extraction. Technical Report DISI-09-055, DISI, University of Trento, Italy (2009)
30.Medelyan, O., Witten, I. H.: Domain independent automatic keyphrase indexing with small training sets. Journal of American Society for Information Science and Technology. Vol. 59 (7), pp. 1026-1040 (2008)
31.Marujo, L., Gershman, A. Carbonell, J., Frederking, R., Neto, J. P.: Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization. Proceedings of LREC 2012 (2012)
32.Marujo, L., Viveiros, M., Neto, J. P.: Keyphrase Cloud Generation of Broadcast News. In proceeding of 12th Annual Conference of the International Speech Communication Association, Interspeech (2011)