Toward Network-based Keyword Extraction from Multitopic Web Documents

Sabina Šišović, Sanda Martinčić-Ipšić, Ana Meštrović
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{ssisovic, smarti, amestrovic}@uniri.hr

Abstract. *In this paper we analyse the selectivity measure calculated from the complex network in the task of the automatic keyword extraction. Texts, collected from different web sources (portals, forums), are represented as directed and weighted co-occurrence complex networks of words. Words are nodes and links are established between two nodes if they are directly co-occurring within a sentence. We test different centrality measures for ranking nodes - keyword candidates. The promising results are achieved using the selectivity measure. Then we propose an approach which enables extracting word pairs according to the values of the in/out-selectivity and weight measures combined with filtering.*

Keywords. keyword extraction, complex networks, co-occurrence language networks, Croatian texts, selectivity

1 Introduction

Keyword extraction is an important task in the domain of the Semantic Web development. It is a problem of automatic identification of the important terms or phrases in text documents. It has numerous applications: information retrieval, automatic indexing, text summarization, semantic description and classification, etc. In the case of web documents it is a very demanding task: it requires extraction of keywords from web pages that are typically noisy, overburden with information irrelevant to the main topic (navigational information, comments, future announcements, etc.) and they usually contain several topics [3]. Therefore, in keyword extraction from web pages we are dealing with noisy and multitopic datasets.

Various approaches have been proposed for keywords and keyphrases identification (extraction) task. There are two main classes of approaches: supervised and unsupervised. Supervised approaches are based on using machine learning techniques on the manually annotated data [19, 20]. Therefore supervised approaches are time consuming and expensive. Unsupervised approaches may include clustering [7], language modelling [18] and graph-based approaches. Unsupervised approaches may also require different sets of

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

external data, however these approaches are not depended on manual annotation. These approaches are more robust, but usually less precise [2].

A class of graph-based keyword extraction algorithms overcome some of these problems. In graph-based or network-based approaches the text is represented as a network in a way that words are represented as nodes and links are established between two nodes if they are co-occurring within the sentence. The main idea is to use different centrality measures for ranking nodes in the network. Nodes with the highest rank represent words that are candidates for keywords and keyphrases. In [5] an exhaustive overview of network centrality measures usage in the keyword identification task is given.

One of the probably most influential graph-based approaches is the TextRank ranking model introduced by Mihalcea and Tarau in [14]. TextRank is a modification of PageRank algorithm and the basic idea of this ranking technique is to determine the importance of a node according to the importance of its neighbours, using global information recursively drawn from the entire network. However, some recent researches have shown that even simpler centrality measures can give satisfactory results. Boudin [2] and Lahiri et al. [5] compare different centrality measures for keyword extraction task. Litvak and Last [6] compare supervised and unsupervised approach for keywords identification in the task of extractive summarization.

We have already experimented with graph-based approaches for Croatian texts representation. In [12, 13] we described graph-based word extraction and representation from the Croatian dictionary. We used lattice to represent different semantic relations (partial semantic overlapping, more specific, etc.) between words from the dictionary. In [8, 10, 17] we described and analysed network-based representation of Croatian texts. In [10] our results showed that in-selectivity and out-selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. It seems that selectivity measure is able to capture typical word phrases and collocations which are lost during the shuffling procedure. The same holds for English where Masucci and Rodgers [11] found that selectivity somehow captures the specialized local structures in nodes' neighborhood and forms of the morphological structures in text. According to these results, we expected that node selectivity may be potentially important for the text categories differentiation and include it in the set of standard network measures. In [17] we show that the node selectivity measure can capture structural differences between two genres of text.

This was the motivation for further exploration of selectivity for keyword extraction task from Croatian multitopic web documents. We have already analysed the selectivity-based keyword extraction in Croatian news [1]. In this paper we propose an in/out-selectivity based approach combined with filtering to extract keyword candidates from the co-occurrence complex network of text. We design selectivity-based approach as unsupervised, data and domain independent. In its basic form, only the stopwords list is a prerequisite for applying stopwords-filter. As designed, it is a very simple and robust approach appropriate for extraction from large multitopic and noisy datasets.

In Section 2 we present measures for the network structure analysis. In Section 3 we describe datasets and the construction of co-occurrence networks from used text collection. In Section 4 are the results of keyword extraction, and in the final Section 5, we elaborate the obtained results and make conclusions regarding future work.

2 The network measures

This section describes basic network measures that are necessary for understanding our approach. More details about these measures can be found in [11, 15, 16]. In the network, N is the number of nodes and K is the number of links. In weighted language networks every link connecting two nodes i and j has an associated weight w_{ij} that is a positive integer number.

The node degree k_i is defined as the number of links incident upon a node. The in degree and out degree $k_i^{in/out}$ of node i is defined as the number of its in and out neighbours.

Degree centrality of the node i is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$\mathrm{dc}_i = \frac{k_i}{N-1}. \tag{1}$$

Analogously, the in-degree centralities are defined as in-degree of a node:

$$\mathrm{dc}_i^{in} = \frac{k_i^{in}}{N-1}. \tag{2}$$

The out-degree centrality of a node is defined in a similar way. Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest paths between a node and all the other nodes. Let d_{ij} be the shortest path between nodes i and j. The normalised closeness centrality of a node i is given by:

$$\mathrm{cc}_i = \frac{N-1}{\sum_{i \neq j} d_{ij}}. \tag{3}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that traverse through the node i. The normalised betweenness centrality of a node i is given by:

$$\mathrm{bc}_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{4}$$

The strength of a node i is a sum of weights of all links incident with the node i:

$$\mathrm{s}_i = \sum_j w_{ij}. \tag{5}$$

All given measures are defined for directed networks, but language networks are weighted, therefore, the weights should be considered. In the directed network, the in-strength s_i^{in} of the node i is defined as the number of its incoming links, that is:

$$\mathrm{s}_i^{in} = \sum_j w_{ji}. \tag{6}$$

The out-strength is defined in a similar way. The selectivity measure is introduced in [11]. It is actually an average strength of a node. For a node i the selectivity is calculated as a fraction of the node weight and node degree:

$$e_i = \frac{s_i}{k_i}. \quad (7)$$

In the directed network, the in-selectivity of the node i is defined as:

$$e_i^{in} = \frac{s_i^{in}}{k_i^{in}}. \quad (8)$$

The out-selectivity is defined in a similar way.

3 Methodology

3.1 The construction of co-occurrence networks

Dataset contains 4 collections of web documents written in Croatian language collected from different web sources (portals and forums on different daily topics). The 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).

The first step in networks construction was text preprocessing: "cleaning" special symbols, normalising Croatian diacritics (č, ć, ž, š, dž), and removing punctuations which does not mark the end of a sentence. Commonly, for Croatian which is highly flective Slavic language the lemmatisation and part-of-speech tagging should be performed, but we model our experiment without any explicit language knowledge.

For each dataset we constructed weighted and directed co-occurrence network. Nodes are words that are linked if they are direct neighbours in a sentence. The next step was introducing the networks as weighted edgelists, which contain all the pairs of connected words and their weights (the number of connections between two same words). In the Table 1 there are number of words, number of nodes and number of links per each dataset. We used Python and the NetworkX software package developed for the construction, manipulation, and study of the structure, dynamics, and functions of complex networks [4].

3.2 The selectivity-based approach

The goal of this experiment is to analyse the selectivity measure in the automatic keyword extraction task. First, we compute centrality measures for each node in all 4 networks: in-degree centrality, out-degree centrality, closeness centrality, betweenness centrality and selectivity centrality. Then we rank all nodes (words) according to the values of each of these measures, obtaining top 10 keyword candidates automatically from the network.

In the second part of our experiment we compute in-selectivity and out-selectivity for each node in all 4 networks. The nodes are then ranked according to the highest in/out-selectivity values. Then, for every node we detect neighbour nodes with the highest weight. For the in-selectivity we isolate one neighbour node with the highest outgoing link weight. For the out-selectivity we isolate one neighbour node with the highest ingoing link weight. The result of in/out-selectivity extraction is a set of ranked word tuples.

The third part of our approach consider applying different filters on the in/out-selectivity based word tuples. The first is the stopwords-filter: we filter out all tuples that contain stopwords. Stopwords are a list of the most common, short function words which

Dataset	GL	NN	IN	SD
Number of words	199 417	146 731	118 548	44 367
Number of nodes N	27727	13036	15065	9553
Number of links K	105171	55661	28972	25155

Table 1: The number of words, number of nodes and number of links for all 4 datasets

	selectivity	in-degree	out-degree	closeness	betweenness
1.	mladićevi (joungsters)	i (and)	i (and)	je (is)	i (in)
2.	pomlatili (beaten)	u (in)	je (is)	i (and)	je (is)
3.	seksualnog (sexual)	je (is)	u (in)	se (self)	u (in)
4.	policijom (police)	na (on)	na (on)	da (that)	na (on)
5.	uhićeno (arrested)	da (that)	se (self)	su (are)	se (self)
6.	skandala (scandal)	za (for)	za (for)	to (it)	za (for)
7.	podnio (submitted)	se (self)	su (are)	a (but)	da (that)
8.	obožavatelji (fans)	a (but)	da (that)	će (will)	su (are)
9.	sata (hour)	su (are)	s (with)	samo (only)	a (but)
10.	Baskiji (Baskia)	s (with)	od (from)	ali (but)	s (with)

Table 2: Top ten words from the dataset IN ranked according to the selectivity, in/out-degree, closeness and betwenness

do not carry strong semantic properties, but are needed for the syntax of language (pronouns, prepositions, conjunctions, abbreviations, interjections,...). The second is the high-weights-filter: from the in/out-selectivity based word tuples we chose only those tuples that have the same values for the selectivity and weight. The third filter is the combination of the first two filters.

4 Results

Initially, we analyse 4 networks constructed for each dataset. The top 10 ranked nodes with the highest values of the selectivity, in degree, out degree, closeness and betwenness measures for datasets IN, GL, SD and NN are shown in the Tables 2,3,4 and 5. It is obvious that top 10 ranked words according to the in/out degree centrality, closeness centrality and betwenness centrality are stopwords. It can be also noticed that centrality measures return almost identical top 10 stopwords. However, the selectivity measure ranked only open-class words: nouns, verbs and adjectives. We expect that among these highly ranked words are keyword candidates.

Furthermore, we analyse selectivity measure in details. Since texts are better represented as directed networks [9], we analyse words with in-selectivity and out-selectivity measure separately. We extract word-tuple: the word before for in-selectivity and the word after for out-selectivity that has the highest value of the weight. In Table 6 are shown ten highly ranked in/out-selectivity based word-tuples together with the values of in/out-selectivity and weight.

Hence, we extract the most frequent word-tuples which are possible collocations or phrases from the text. We expect that among these highly ranked word-tuples are keyword

	selectivity	in degree	out degree	closeness	betweenness
1.	stupastih (cage)	i (and)	i (and)	i (and)	i (and)
2.	populaciju (population)	u (in)	u (in)	se (self)	u (in)
3.	izdanje (issue)	na (on)	je (is)	je (is)	je (is)
4.	online (online)	je (is)	se (self)	su (are)	na (on)
5.	webshop (webshop)	ili (or)	na (on)	a (but)	se (self)
6.	matrica (matrix)	a (but)	ili (or)	ili (or)	ili (or)
7.	pretplata (subscription)	se (self)	su (are)	to (it)	a (but)
8.	časopis (journal)	za (for)	za (for)	bolesti (disease)	za (for)
9.	oglasi (ads)	od (from)	od (from)	da (that)	su (are)
10.	marketing (marketing)	su (are)	a (but)	biljke (plants)	od (from)

Table 3: Top ten words from the dataset GL ranked according to the selectivity, in/out-degree, closeness and betwenness

	selectivity	in-degree	out-degree	closeness	betweenness
1.	seronjo (bullshitter)	i (and)	i (and)	i (and)	i (and)
2.	Splitu (Split)	u (in)	je (is)	je (is)	je (is)
3.	upišite (fill-in)	je (is)	u (in)	svibanj (May)	u (in)
4.	uredniku (editor)	komentar (comment)	se (self)	se (self)	se (self)
5.	ekrana (monitor)	na (on)	svibanj	ali (but)	na (on)
6.	crkvu (church)	se (self)	na (on)	a (but)	od (from)
7.	supetarski (Supetar)	za (for)	za (for)	će (will)	za (for)
8.	vijesti (news)	a (but)	da (that)	to (it)	a (but)
9.	zaradom (earning)	svibanj (May)	ne (ne)	još (more)	svibanj
10.	Jović (Jović)	od (from)	a (but)	pa (so)	to (it)

Table 4: Top ten words from the dataset SD ranked according to the selectivity, in/out-degree, closeness and betwenness

	selectivity	in-degree	out-degree	closeness	betweenness
1.	novine (newspaper)	i (and)	i (and)	i (and)	i (and)
2.	temelju (based on)	u (in)	u (in)	ili (or)	u (in)
3.	manjinu (minority)	za (for)	je (is)	je (is)	za (for)
4.	srpsku (Serbian)	na (on)	za (for)	se (self)	ili (or)
5.	sladu (harmony)	ili (or)	se (self)	da (that)	na (on)
6.	snagu (strength)	iz (from)	ili (or)	usluga (service)	je (is)
7.	osiguranju (insurance)	te (and)	na (on)	zakona (law)	se (self)
8.	narodnim (national)	je (is)	o (on)	a (but)	o (on)
9.	novinama (newspaper)	se (self)	te (and)	skrbi (welfare)	te (and)
10.	kriza (crisis)	s (with)	članak (article)	HRT-a (HRT-a)	iz (form)

Table 5: Top ten words from the dataset NN ranked according to the selectivity, in/out-degree, closeness and betwenness

	in-selectivity			out-selectivity		
	word tuple	e^{in}	w	word tuple	e^{out}	w
1.	narodne **novine**	326	326	**srpsku** nacionalnu	222	222
2.	na **temelju**	317	317	**nacionalnu** pripadnost	183	1
3.	nacionalnu **manjinu**	275	2	**ovjesne** jedrilice	159	159
4.	za **srpsku**	222	222	**narodnim** novinama	129	129
5.	u **skladu**	202	202	**narodne** jazz	111	1
6.	na **snagu**	172	172	**manjinu** gradu	78	1
7.	o **osiguranju**	134	43	**ovoga** sporazuma	72	1
8.	u **narodnim**	129	129	**crvenog** kristala	72	3
9.	narodnim **novinama**	129	129	**skladu** provjeriti	67	1
10.	crvenog **križa**	99	2	**oružani**h sukoba	58	4

Table 6: Top ten highly ranked in/out-selectivity based word-tuples for the NN dataset

	in-selectivity			out-selectivity		
	word tuple	e^{in}	w	word tuple	e^{out}	w
1.	narodne **novine**	326	326	**srpsku** nacionalnu	222	222
2.	nacionalnu **manjinu**	275	2	**nacionalnu** pripadnost	183	1
3.	narodnim **novinama**	129	129	**ovjesne** jedrilice	183	1
4.	crvenoga **križa**	99	2	**narodnim** novinama	129	129
5.	jedinicama **regionalne**	65	1	**narodne** jazz	111	1
6.	nacionalne **manjine**	61	61	**manjinu** gradu	78	1
7.	rizika **snaga**	57	1	**ovoga** sporazuma	72	1
8.	medije **ubroj**	47	1	**crvenog** kristala	72	3
9.	crveni **križ**	42	42	**skladu** provjeriti	67	1
10.	uopravni **spor**	41	41	**oružanih** sukoba	58	4

Table 7: Top ten highly ranked in/out-selectivity based word-tuples without stopwords for the NN dataset

in-selectivity		out-selectivity	
word tuple	e^{in}=w	word tuple	e^{out}=w
na **temelju** (based on)	317	**ovjesne** jedrilice (hangh glider)	159
za **srpsku** (for Serbian)	222	**narodnim** novinama (Nat. news.)	129
u **skladu** (according to)	202	**sjedištem** u (headquarter in)	55
na **snagu** (into effect)	172	**objavit** će (will be bublished)	53
u **narodnim** (in national)	129	**republici** Hrvatskoj (Croatia)	52
narodnim **novinama** (Nat. news.)	129	**albansku** nacionalnu (Alb. nat.)	52
i **dopunama** (and amendments)	68	**republika** Hrvatska (Croatia)	49
nacionalne **manjine** (nat. minority)	61	**oplemenjivačkog** prava (noble law)	45
sa **sjedištem** (with headquarter)	55	**madjarsku** nacionalnu (Hung. nat.)	40

Table 8: Top ten highly ranked in/out-selectivity based word-tuples with equal in/out-selectivity and weight for the NN dataset

in-selectivity word tuple	out-selectivity word tuple
narodne **novine** (National newspaper)	**srpsku** nacionalnu (Serbian national)
narodnim **novinama** (Nat. newspapers)	**ovjesne** jedrilice (hangh glider)
nacionalne **manjine** (nat. minority)	**narodnim** novinama (Nat. newspapers)
crveni **križ** (red cross)	**republici** hrvatskoj (Republic of Croatia)
upravni **spor** (administrative dispute)	**albansku** nacionalnu (Albanian national)
ovjesnom **jedrilicom** (hangh glider)	**republika** hrvatska (Republic of Croatia)
elektroničke **medije** (electronic media)	**oplemenjivačkog** prava (noble law)
nacionalnih **manjina** (national minority)	**madjarsku** nacionalnu (Hungarian nat.)
domovinskog **rata** (Homeland War)	**romsku** nacionalnu (Romany national)
Ivan **Vrljić** (Ivan Vrljić)	**nadzorni** odbor (supervisory board)

Table 9: Top ten highly ranked in/out-selectivity based word-tuples with equal in/out-selectivity and weight without stopwords for the NN dataset

candidates. Due to limited space, we show results only for the NN dataset, but other datasets raised similar results.

In Table 6 there are word-tuples which contain stopwords, especially for the in-selectivity based ranking.Therefore we use stopwords-filter defined in the previous section as shown in Table 7. Now we obtain more open class keyword candidates from highly ranked word-tuples.

In Table 8. there are 10 highly ranked word-tuples for the NN dataset with the high-weights-filter applied. Using this approach some new keyword candidates appear in the ranking results.

In Table 9. there are 10 highly ranked word-tuples from the NN dataset with the both filters applied. According to our knowledge about the content of the dataset, these two filters derived the best results.

5 Conclusion and discussion

We analyse network-based keyword extraction from multitopic Croatian web documents using selectivity measure. We compare keyword candidate words rankings with selectivity and three network-based centrality measures (degree, closeness and betwenness). The selectivity measure gives better results because centrality-based rankings select only stopwords as the top 10 ranked words. Furthermore, we propose extracting the highly connected word-tuples with the highest in/out-selectivity values as the keyword candidates. Finally, we apply different filters (stopwords-filter, high-weights-filter) in order to keyword candidate list.

The first part of analysis can raise some considerations regarding the selectivity measure. The selectivity measure is important for the language networks especially because it can differentiate between two types of nodes with high strength values (which means words with high frequencies). Nodes with high strength values and high degree values would have low selectivity values. These nodes are usually stopwords (conjunctions, prepositions,...). On the other side, nodes with high strength values and low degree values would have high selectivity values. These nodes are possible collocations, keyphrases and names that appear in the texts. It seems that selectivity is insensitive to stopwords (which are the most frequent words) and therefore can efficiently detect semantically rich open class words from the network.

Furthermore, since we modelled multitopic datasets the keyword extraction task is even more demanding. From the results of this preliminary research it seems that the selectivity has a potential to extract keyword candidates without preprocessing (lemmatisation, POS tagging) from multitopic sources.

There are several drawbacks in this reported work: we did not perform the classical evaluation procedure because we did not have annotated data and we conducted analysis only on Croatian texts.

For the future work we plan to evaluate our results on different datasets in different languages. Furthermore, it seems promising to define an approach that can extract a sequence of three or four neighbouring words based on filtered word-tuples. We also plan to experiment with lemmatised texts. Finally, in the future we will examine the effect of noise to the results obtained from multitopic sources.

References

[1] S. Beliga, A. Meštrović and S. Martinčić-Ipšić. Toward Selectivity Based Keyword Extraction for Croatian News. Submitted on Workshop on Surfacing the Deep and the Social Web, Co-organised by ICT-COST Action KEYSTONE (IC1302), Riva Del Garda, Trento, Italy, 2014.

[2] F. Boudin. A comparison of centrality measures for graph-based keyphrase extraction. International Joint Conference on Natural Language Processing. pp. 834–838, (2013)

[3] M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multi-theme documents. ACM 18th conference on World wide web, pp.661–670, (2009)

[4] A. Hagberg, P. Swart, and D. Chult. Exploring network structure, dynamics, and function using networkx. (2008)

[5] S. Lahiri, S.R. Choudhury, and C. Caragea. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks. arXiv:1401.6571, (2014)

[6] M. Litvak and M. Last. Graph-based keyword extraction for single-document summarization. ACM Workshop on Multi-source Multilingual Information Extraction and Summarization. pp.17–24, (2008)

[7] Z. Liu, P. Li, Y. Zheng, and M. Sun. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Volume 1-Volume 1, pp. 257-266 (2009)

[8] D. Margan, S. Martinčić-Ipšić and A. Meštrović. Network Differences Between Normal and Shuffled Texts: Case of Croatian. Studies in Computational Intelligence, Complex Networks V. Vol.549. Italy, pp. 275–283 (2014)

[9] D. Margan, S. Martinčić-Ipšić, and A. Meštrović. Preliminary report on the structure of Croatian linguistic co-occurrence networks. 5th International Conference on Information Technologies and Information Society (ITIS), Slovenia, 89–96 (2013)

[10] D. Margan, A. Meštrović and S. Martinčić-Ipšić. Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts. IEEE MIPRO 2014, Croatia, pp.1819–1823 (2014)

[11] A. Masucci and G. Rodgers. Differences between normal and shuffled texts: structural properties of weighted networks. Advances in Complex Systems, 12(01):113–129 (2009)

[12] A. Meštrović and M. Čubrilo. Monolingual dictionary semantic capturing using concept lattice. International Review on Computers and Software 6(2):173–184 (2011)

[13] A. Meštrović. Semantic Matching Using Concept Lattice. In Proc. of Concept Discovery in Unstructured Data, Katholieke Universiteit Leuven, pp. 49–58 (2012)

[14] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. ACL Empirical Methods in Natural Language Processing, (2004)

[15] M. E. J. Newman. Networks: An Introduction. Oxford University Press.(2010)

[16] T. Opsahl, F. Agneessens and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3): 245–251 (2010)

[17] S. Šišović, S. Martinčić-Ipšić and A. Meštrović. Comparison of the language networks from literature and blogs. IEEE MIPRO 2014, Croatia, pp.1824–1829, (2014).

[18] T. Tomokiyo, and M. Hurst. A language model approach to keyphrase extraction. In Proc. of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment,Volume 18, pp. 33–40 (2003)

[19] P. D. Turney. Learning algorithms for keyphrase extraction. 2(4):303–336 (2000)

[20] I. H. Witten et al. Nevill-Manning. Kea: practical automatic keyphrase extraction. In Proc. of the fourth ACM conference on Digital libraries, pp. 254-255 (1999)