Toward Selectivity-Based Keyword Extraction for Croatian News

Slobodan Beliga, Ana Meštrović, Sanda Martinčić-Ipšić

Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
{sbeliga,amestrovic,smarti}@uniri.hr

Abstract. Our approach proposes a novel network measure - the node selectivity for the task of keyword extraction. The node selectivity is defined as the average strength of the node. Firstly, we show that selectivity-based keyword extraction slightly outperforms the extraction based on the standard centrality measures: in-degree, out-degree, betweenness, and closeness. Furthermore, from the data set of Croatian news we extract keyword candidates and expand extracted nodes to word-tuples ranked with the highest in/out selectivity values. The obtained sets are evaluated on manually annotated keywords: for the set of extracted keyword candidates the average $F1$ score is 24.63%, and the average $F2$ score is 21.19%; for the exacted word-tuples candidates the average $F1$ score is 25.9% and the average $F2$ score is 24.47%. Selectivity-based extraction does not require linguistic knowledge while it is purely derived from statistical and structural information of the network.

Keywords: keyword extraction, complex network, centrality measures, selectivity, Croatian news texts

1 Introduction

The task of keyword extraction is to automatically identify a set of terms that best describe the document [14]. Automatic keyword extraction establishes a foundation for various natural language processing applications: information retrieval, the automatic indexing and classification of documents, automatic summarization, high-level semantic description, etc.

State-of-the-art keyword extraction approaches are based on statistical methods which require learning from hand-annotated data sets. In the last decade the focus of research has shifted toward unsupervised methods, mainly towards network or graph enabled keyword extraction. In a network enabled keyword extraction the document representation may vary from very simple (words are nodes and their co-occurrence is represented with links), or can incorporate very sophisticated linguistic knowledge like syntactic [2] or semantic relations [18]. Typically, the source (document, text, data) for keyword extraction is modelled with one network. This way, both the statistical properties (frequencies) as well as the structure of the source text are represented by a unique formal representation, hence a complex network.

A network (or graph, since the number of words in isolated documents is limited) enabled keyword extraction exploits different measures for the task of identifying and ranking the most representative features of the source - the keywords. The keyword extraction powered by network measures can be on the node, network or subnetwork level. Measures on the node level are: degree, strength, centrality [9]; on the network level: coreness, clustering coefficient, PageRank motivated ranking score or HITS motivated hub and authority score [10, 11, 14]; on the subnetwork level: communities [12]. Most of the of the research was motivated with various centrality measures: degree, betweenness, closeness and eigenvector centrality [9–11, 13–15].

Our research aims at proposing a novel selectivity-based method for the unsupervised keyword extraction from the network of Croatian texts. Since Croatian is a highly flective Slavic language, the source text usually needs a substantial preprocessing (lemmatization - morphological normalization, stopwords removal, part-of-speech (POS) annotation, morphosyntactic descriptions (MSD) tagging, etc.), we design our approach with little or no linguistic knowledge. A new network measure - the node selectivity, originally proposed by Masucci and Rodgers [7, 8] (that can distinguish a real from a shuffled one), is applied to automatic keyword extraction. Selectivity is defined as the average weight distribution on the links of the single node. In our previous work, the node selectivity measure performed in favour of the differentiation between original and shuffled Croatian texts [4, 5], and for the differentiation of blog and literature text genres [6]. In this work we explore the potential of the selectivity for the keyword extraction in the Croatian news articles. To the best of our knowledge, the node selectivity measure has not been applied to the keyword extraction task before.

Section 2 presents an overview of related work on automatic keyword extraction. In Section 3 we present the definition of the measures for the network structure analysis. In Section 4 we present the construction of co-occurrence networks from collection of used text. The methods used for network based keyword extraction are explained in Section 5. The evaluation of obtained keywords and results are in Section 6. In the final Section, we elaborate upon the selectivity method and make conclusions regarding future work.

2 Related Work

Lahiri et al. [9] extract keywords and keyphrases form co-occurrence networks of words and from noun phrases collocations networks. Eleven measures (degree, strength, neighbourhood size, coreness, clustering coefficient, structural diversity index, page rank, HITS hub and authority score, betweenness, closeness and eigenvector centrality) are used for keyword extraction from directed/undirected and weighted networks. The obtained results on 4 data sets suggest that centrality measures outperform the baseline term frequency/inverse document frequency (tf-idf) model, and simpler measures like degree and strength outperform computationally more expensive centrality measures like coreness and betweenness.

Boudin [10] compares various centrality measures for graph-based keyphrase extraction. Experiments on standard data sets of English and French show that simple degree centrality achieves results comparable to the widely used TextRank algorithm; and that closeness centrality obtains the best results on short documents. Undirected and weighted co-occurrence networks are constructed from syntactically (only nouns and adjectives) parsed and lemmatized text using co-occurrence window. Degree, closeness, betweenness and eigenvector centrality are compared to PageRank ad proposed by Mihalcea in [14] as a baseline. Degree centrality achieve similar performance as much complex TextRank. Closeness centrality outperforms TextRank on short documents (scientific papers abstracts).

Litvak and Last [11] compare supervised and unsupervised approaches for keywords identification in the task of extractive summarization. The approaches are based on the graph-based syntactic representation of text and web documents. The results of the HITS algorithm on a set of summarized documents performed comparably to supervised methods (Naive Bayes, J48, Support Vector Machines). The authors suggest that simple degree-based rankings from the first iteration of HITS, rather than running it to its convergence, should be considered.

Grineva et al. [12] use community detection techniques for key terms extraction on Wikipedia's texts, modelled as a graph of semantic relationships between terms. The results showed that the terms related to the main topics of the document tend to form a community, thematically cohesive groups of terms. Community detection allows the effective processing of multiple topics in a document and efficiently filters out noise. The results achieved on weighted and directed networks from semantically linked, morphologically expanded and disambiguated n-grams from the article's titles. Additionally, for the purpose of the noise stability, they repeated the experiment on different multi-topic web pages (news, blogs, forums, social networks, product reviews) which confirmed that community detection outperforms td-idf model.

Palshikar [13] proposes a hybrid structural and statistical approach to extract keywords from a single document. The undirected co-occurrence network, using a dissimilarity measure between two words, calculated from the frequency of their co-occurrence in the preprocessed and lemmatized document, as the edge weight, was shown to be appropriate for the centrality measures based approach for keyword extraction.

Mihalcea and Tarau [14] report a seminal research which introduced a state-of-the-art TextRank model. TextRank is derived from PageRank and introduced to graph based text processing, keyword and sentence extraction. The abstracts are modelled as undirected or directed and weighted co-occurrence networks using a co-occurrence window of variable sizes (2..10). Lexical units are preprocessed: stopwords removed, words restricted with POS syntactic filters (open class words, nouns and adjectives, nouns). The PageRank motivated score of the importance of the node derived from the importance of the neighboring nodes is used for keyword extraction. The obtained TextRank performance compares favorably with the supervised machine learning n-gram based approach.

Matsou et al. in [15] present an early research where a text document is represented as an undirected and unweighted co-occurrence network. Based on the network topology, the authors proposed an indexing system called KeyWorld, which extracts important terms (pairs of words) by measuring their contribution to small-world properties. The contribution of the node is based on closeness centrality calculated as the difference in small-world properties of the network with the temporarily elimination of a node combined with inverse document frequency (idf).

Erkan and Radev [16] introduce a stochastic graph-based method for computing the relative importance of textual units on the problem of text summarization by extracting the most important sentences. LexRank calculates sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. LexRank is shown to be quite insensitive to the noise in the data.

Mihalcea in [17] presents an extension to earlier work [14], where the TextRank algorithm is applied for the text summarization task powered by sentence extraction. On this task TextRank performed on a par with the supervised and unsupervised summarization methods, which motivated the new branch of research based on the graph-based extracting and ranking algorithms.

Tsatsaronis et al. [18] present SemanticRank, a network based ranking algorithm for keyword and sentence extraction from text. Semantic relation is based on the calculated knowledge-based measure of semantic relatedness between linguistic units (keywords or sentences). The keyword extraction from the Inspec abstracts' results reported a favorable performance of SemanticRank over state-of-the-art counterparts - weighted and unweighted variations of PageRank and HITS.

Huang et al. [19] propose an automatic keyphrase extraction algorithm using an unsupervised method based on connectedness and betweeness centrality.

2.1 Related Work on the Croatian Language

The keyphrase extraction for the Croatian language has been addressed in both supervised [23] and unsupervised [20–22] settings. Ahel et al. [23] use a Naive Bayes classifier combined with tf-idf (term frequency/inverse document frequency), [20] utilizes the part-of-speech (POS) and morphosyntactic description (MSD) tags filtering followed by tf-idf ranking, and [22] exploits the distributional semantics to build topically related word clusters, from which they extract keywords and expand them to keyphrases. Bekavac et al. [21] propose a genetic programming approach for keyphrases the extraction for the Croatian language on the same data set. GPKEX can evolve simple and interpretable keyphrase scoring measures that perform comparably to other machine learning methods for Croatian. Reported research on extraction of Croatian keywords use a data set composed of Croatian news articles from the Croatian News Agency (HINA), with hand annotated keywords by human experts.

3 The Complex Network Analysis

This section describes the basic network measures that are necessary for understanding our approach. More details about these measures can be found in [8, 24]. 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} which is a positive integer number.

The node degree k_i is defined as the number of edges 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 $N-1$:

$$dc_i = \frac{k_i}{N-1}. \tag{1}$$

Analogue, the in/out degree centralities are defined as in/out degree of a node:

$$dc_i^{(in/out)} = \frac{k_i^{(in/out)}}{N-1}. \tag{2}$$

Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest distances 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:

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

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 the shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that pass through the node i. The normalised betweenness centrality of a node i is given by:

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

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

$$s_i = \sum_j w_{ij}. \tag{5}$$

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/out strength $s_i^{in/out}$ of the node i is defined as the number of its incoming and outgoing links, that is:

$$s_i^{in/out} = \sum_j w_{ji/ij}. \tag{6}$$

The selectivity measure is introduced in [8]. It is actually an average strength of a node. For the node i the selectivity is calculated as a fraction of the node weight and node degree:

$$e_i = \frac{s_i}{k_i}. \tag{7}$$

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

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}. \tag{8}$$

4 Methodology

4.1 Data

For the network based keyword extraction we use the data set composed of Croatian news articles [20]. The data set contains 1020 news articles from the Croatian News Agency (HINA), with manually annotated keywords (key phrases) by human experts. The set is divided: 960 annotated documents for learning of supervised methods, and 60 documents for testing. The test set of 60 documents is annotated by 8 different experts, where the inter-annotator agreement in terms of F2 scores (see Section 5) are on average 46% (between 29.3% and 66.1%).

We selected the first 30 texts from the HINA collection for our experiment. The texts required some preprocessing: parsing only textual part and title part excluding annotations, cleaning of diacritics and symbols (w instead of vv, ! instead of l, etc.) and lemmatization. Non-standard word forms numbers, dates, acronyms, abbreviations etc. remain in text, since the method is preferably resistant to the noise presented in the data source.

The selected 30 texts varied in length: from very short 60 tokens up to 800 tokens (318 tokens on average). The number of keywords per document varies between 9 and 42 (24 on average). One annotator on average annotated 10 keywords per document.

4.2 The Construction of Co-occurrence Networks

Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. Co-occurrence networks exploit simple neighbour relation, two words are linked if they are adjacent in the sentence [3]. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus.

From the documents in the HINA data set we construct directed and weighted co-occurence networks: one from the text in each document and an integral one from the texts in all documents; 31 in total.

Network construction and analysis was implemented with the Python programming language using the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics and functions of complex networks [1].

5 Keyword Extraction

5.1 Centrality Motivated Keyword Extraction

Network based keyword extraction methods exploit different measures for the task of identification and ranking the most representative features of the source - the keywords. The first part of our research compares the performance of different centrality motivated network measures (in/out degree, closeness and betweenness) with the performance of proposed selectivity measure. The second part develops a selectivity-based method for keyword extraction with a comparative analysis of unsupervised (non-network enabled) approaches.

The degree (Eq. 1 and 2) of a node (word) is the number of neighbouring nodes (different neighbouring words). Typically, the nodes with the highest degree in the network are hubs, analogously the words with the highest degree are expectedly stopwords. The closeness (Eq. 3) of a node (word) is related to the farness of the word from all other words in the text. The betweenness (Eq. 4) of a node (word) is the measure of how many shortest paths between all other node-pairs are traversing a node. The words with the highest values of the betweenness centrality are considered to be important for the information flow as well. Selectivity is a local (node level) network measure, defined as the ratio of the node strength and the node degree. In weighted and directed co-occurrence networks one can consider the in- and out- links for obtaining in/out selectivity of the node (Eq. 8). The computation of the node's selectivity value is less complex and expensive than the computation of closeness and betweenness values.

From the network constructed from all the texts in the HINA news data set we calculate in/out degree, closeness, betweenness and in/out selectivity. Based on the obtained values we rank the top 10 or the top 24 keyword candidates from the network and evaluate them on the set of manually annotated keywords, as presented in Table 1. The top 10 or the top 24 keywords are selected due to the average number of human assigned keywords: on average 10 keywords from one annotator and on average 24 keywords from all 8 annotators per document. We evaluate the performance of each network measure based on standard recall (R), precision (P) and $F1$ score. $F1$ score is a harmonic mean of precision and recall: $F_1 = 2PR/(P + R)$. Beside the standard $F1$ score we also calculate the $F2$ score, which gives twice as much importance to the recall as to the precision: $F_2 = 5PR/(4P + R)$.

Table 1. The top 10 and top 24 highly ranked keyword candidates form in-degree, out-degree, closeness, betweenness and in/out selectivity values obtained from all the HINA texts' network in terms of Recall (R), Precision (P), $F1$ and $F2$ score

	TOP 10				TOP 24			
	R[%]	P[%]	$F1$[%]	$F2$[%]	R[%]	P[%]	$F1$[%]	$F2$[%]
In-degree	0	0	0	0	0.19	33.33	0.38	0.24
Out-degree	0	0	0	0	0.37	40.00	0.73	0.46
Closeness	0	0	0	0	0.75	**66.67**	1.48	0.93
Betweenness	0.19	**50.00**	0.38	0.24	0.37	50.00	0.73	0.46
In/out selectivity	**0.75**	40.00	**1.47**	**0.93**	**1.31**	29.17	**2.51**	**1.62**

The results in Table 1 are in favour of the selectivity over other standard centrality network measures. The selectivity can efficiently differentiate between two basic types of nodes (words). The nodes with high strength and high degree values, have low selectivity and they are usually closed-class words (e.g. stopwords, conjunctions, prepositions). The nodes with high strength and low degree have high selectivity values. Typically, the highest selectivity value nodes are open-class words which are preferred keyword candidates (nouns, adjectives, verbs) or even part of collocations, keyphrases, names, etc. On the other hand, the highest ranked words with in/out degree, closeness and betweenness are stopwords, which are not suitable keyword candidates. For example the top 10 ranked words according to in-degree centrality are: *to be, and, in, on, which, for, but, this, self, of*; according to betweenness they are: *to be, and, in, on, self, this, which, for, Croatian, but*; according to in/out selectivity they are: *Bratislava, area, Tuesday, inland, revolution, verification, decade, Balkan, freedom, Universe.*

In short, it seems that selectivity is insensitive to stopwords (the most frequent function words, which do not carry strong semantic properties, but are needed for the syntax of language) and therefore can efficiently detect semantically rich open-class words from the network and extract better keyword candidates.

5.2 Selectivity-Based Keyword Extraction

The second part of our research develops a selectivity-based method for keyword extraction. In order to compare the selectivity-based extraction to non-network based approaches (unsupervised machine learning methods) we construct 30 networks (directed and weighted) from the 30 texts in the HINA data set and evaluate with manually annotated keyword sets.

From 30 networks we compute in/out selectivity for all nodes. The nodes are ranked according to the highest in/out selectivity values above a threshold value. Preserving the same threshold value (≥ 1) in all documents resulted in different number of nodes (one word long keyword candidates) extracted per each network. The obtained set of one word long keyword candidates is noted as SET1.

Then, for every filtered node we detect neighbouring nodes: for the in-selectivity we isolate one neighbour node with the highest outgoing weight; for the out-selectivity we isolate one neighbour node with the highest ingoing weight. The result of in/out selectivity extraction is a set of ranked word-tuples - SET2. Word-tuples are two-word long sequences of keyword candidates. From the obtained tuples we filtered out those containing stopwords in order to compare with the manually annotated evaluation set.

6 Evaluation and Results

For the keyword extraction task the strategy "more is better" can be utilized, since there is no objective judgement on keywords. Hence, it is preferable to extract more keywords which makes trade a off between precision and recall of the methods. The second polemic issue of keyword extraction task is: shorter keywords are more general vs. longer which are more accurate. Motivated by these open arguments, and by the approach of other authors, we decided to follow the same principle: to extract as many keyword candidates as possible and evaluate them on the basis of recall (R) and $F2$ score, beside the standard precision (P) and $F1$ score.

Evaluation is the final part of the experiment based on the intersection of the obtained sets SET1 and SET2 of keyword candidates with the union of all 8 annotators keywords. The results in terms of precision and recall are in Figures 1 and 2 respectively, and in terms of $F1$ and $F2$ scores in Figures 3 and 4 respectively. The obtained average $F1$ score for the SET 1 is 24.63%, and the average $F2$ score is 21.19%. The expansion of obtained candidates to SET2 increased the average $F1$ score to 25.9% and $F2$ score to 24.47%.

All supervised and unsupervised methods reported on keyphrases extraction from the HINA data set incorporate the linguistic knowledge (POS, MSD,..) of Croatian. Mijić et al. [20] initially extracted the list of keyword candidates as a comprehensive list of all words without stopwords) which was expanded into longer n-gram sequences up to a length of four. In [20] a keyphrase extraction system developed for a large-scale Croatian news production system the tf-idf ranking model was used to extract n-grams of up to length of four, which were lemmatized, and POS and MSD filtered. For evaluation the manually annotated key phrases from 60 documents were used. The evaluation set was reduced to keywords suggested only by 3 top annotators (having the highest inter-annotator agreement among all 8 annotators). The results indicate that the performance is comparable to that of the human annotators. Ahel et al. [23] for the one-word long keywords reported precision of 22% and recall of 3.4%.

We designed our method purely from statistical and structural information encompassed in the source text which is reflected in the structure of the network. Our method achieved on a SET1 average recall of 19.53% and precision of 39.1%. Expansion to the word-tuples in SET2 increased average recall to 23.87% and decreased precision to 32.23%. The obtained results are comparable to [20] and [23], but with a slightly different evaluation set up.

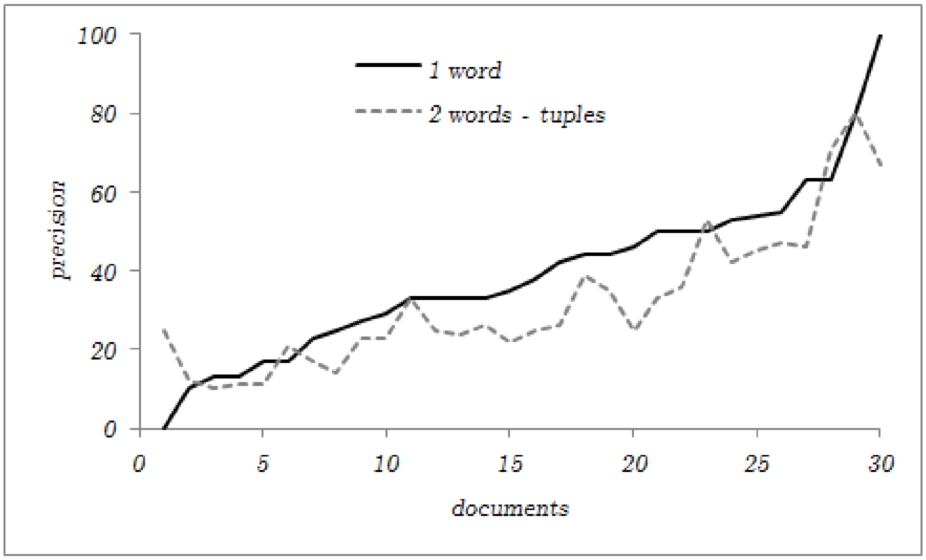

Fig. 1. Precision on the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents

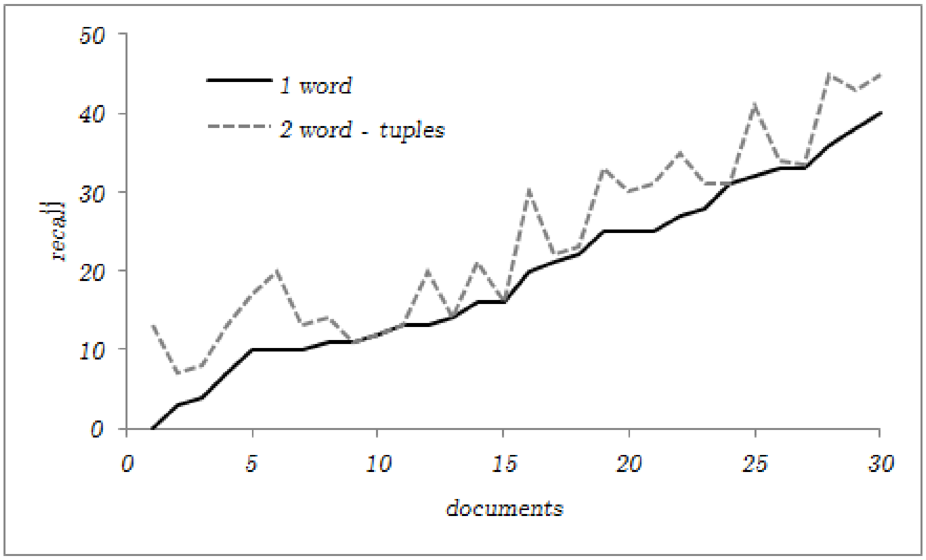

Fig. 2. Recall on the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents

The obtained selectivity-based results are promising and have potential to improve in several directions which is elaborated at the end of the next section. An additional remark regarding results, is that beside keyword candidates our method captures personal names and entities, which were not marked as keyphrases and lowered the score. Capturing names and entities can be of high relevance for the tasks such as name-entity recognition, text summarization, etc.

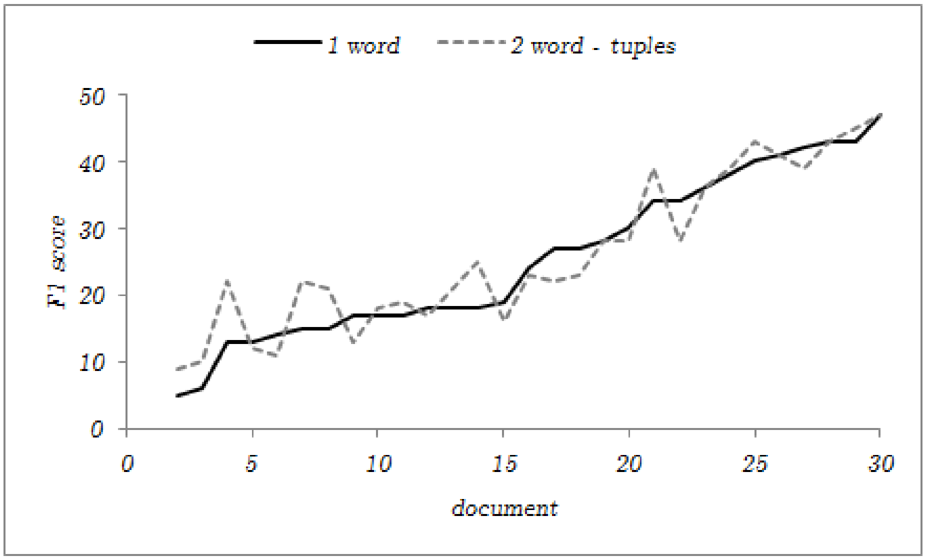

Fig. 3. F1 score of the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents

Keyword annotation is an extremely subjective task as even human experts have difficulties to agree upon keyphrases (inter-agreement around 40%). Croatian is a highly morphologically rich language, which puts another magnitude of challenge on the task, since annotators are freely choosing the morphological word form as a tag, which seems appropriate at the moment. Additionally, there was no predefined set of index or keywords list, so annotators could make up their own, even worse in some cases it seemed appropriate to annotate with keywords, which were not present in the original article (out-of-vocabulray words). In [23] the number of out-of-vocabulary keywords on the whole of the HINA data set is estimated to a high of 57%. Since our method is derived from purely text statistics, it is not capable to capture all the possible subjective variations of the annotators or out-of-vocabulary words. Still it is close to the range of the inter-annotator achieved agreement.

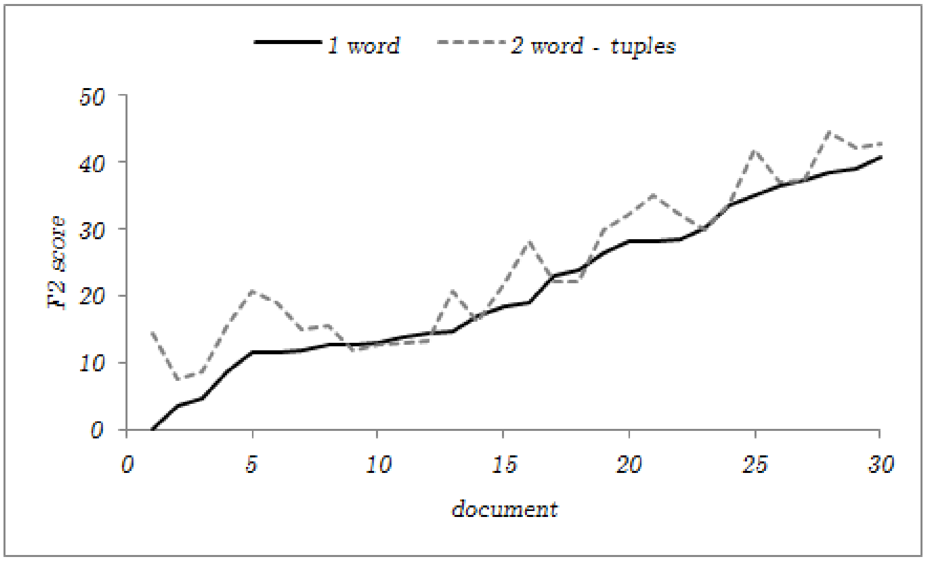

Fig. 4. F2 score of the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents

7 Conclusion

This research on selectivity-based keyword extraction for Croatian news (HINA data set) describes an unsupervised method which extracts nodes from a complex network as keyword candidates. We build our approach with a new network measure - the node selectivity (defined as the average weight distribution on the links of the single node). The node selectivity value is used for extracting and ranking the keyword candidates. Initially, we compare selectivity extraction to standard centrality motivated measures, and propose the selectivity measure for the keyword extraction.

The selectivity-based keyword extraction method is comprised of: the extraction of the seed keyword set (words with the highest in/out selectivity) and expanding them to word-tuples with the highest in/out selectivity values. The obtained average $F1$ score for the set of extracted keyword candidates is 24.63%, and the average $F2$ score is 21.19%. The expansion of the obtained candidates to word-tuples increased the average $F1$ score to 25.9% and $F2$ score to 24.47%, which is comparable to the results on the same data set achieved by supervised and unsupervised methods, and is close to the range of the inter-annotator achieved agreement. The selectivity-based extraction does not require linguistic knowledge as it is purely derived from statistical and structural information encompassed in the source text which is reflected in the structure of the network.

Our results imply that the structure of the network can be applied to the Croatian keyword extraction task with many possible improvements. This should be thoroughly examined in future work, which will cover: a) evaluation - considering all flective word forms; considering various matching strategies - exact, fuzzy, part-of-match; b) text types - considering texts of varying length, genres and topics; c) multitopic - comparing isolate document extraction vs. multitopic extraction; d) other languages - testing on standard English (and other) data sets; e) longer keyword candidate sets - constructing keyword sequences up to a length of 3; f) entity extraction - testing weather entities can be extracted from complex networks.

References

1. A. Hagberg, P. Swart, and D. Chult. Exploring network structure, dynamics, and function using networkX. Technical report, Los Alamos National Laboratory (LANL) (2008)
2. H. Liu and F. Hu. What role does syntax play in a language network? EPL (Europhysics Letters), 83(1):18002 (2008)
3. 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)
4. 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)
5. 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).
6. 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).
7. A. Masucci and G. Rodgers. Network properties of written human language. Physical Review E, 74(2):026102 (2006)
8. 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)
9. S. Lahiri, S.R. Choudhury, and C. Caragea. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks. arXiv preprint arXiv:1401.6571, (2014)
10. F. Boudin. A comparison of centrality measures for graph-based keyphrase extraction. International Joint Conference on Natural Language Processing (IJCNLP), pp. 834–838, (2013)
11. M. Litvak and M. Last. Graph-based keyword extraction for single-document summarization. ACM Workshop on Multi-source Multilingual Information Extraction and Summarization. pp.1724, (2008)
12. M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. ACM 18th conference on World Wide Web, pp.661–670, (2009)
13. G. K. Palshikar. Keyword extraction from a single document using centrality measures. Pattern Recognition and Machine Intelligence, pp.503–510, (2007)
14. R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. ACL Empirical Methods in Natural Language Processing-EMNLP04, (2004)
15. Y. Matsuo, Y. Ohsawa, and M. Ishizuka. Keyworld: Extracting keywords from document s small world. Discovery Science, pp.271–281, (2001)

16. G. Erkan and D. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. Artificial Intelligence Res.(JAIR), vol.22(1), pp.457–479, (2004)
17. R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Proc. ACL 2004, pp.20, (2004)
18. G. Tsatsaronis, I. Varlamis and K. Nørvåg. SemanticRank: ranking keywords and sentences using semantic graphs. ACL 23rd International Conference on Computational Linguistics, pp.1074–1082, (2010)
19. C. Huang, Y. Tian, Z. Zhou, C.X. Ling, and T. Huang. Keyphrase extraction using semantic networks structure analysis., IEEE International Conference on Data Mining, pp.275–284, (2006)
20. J. Mijić, B. Dalbelo-Bašić and J. Šnajder. Robust keyphrase extraction for a large-scale Croatian news production system. FASSBL 2010, pp.59–66, (2010)
21. M.Bekavac and J. Šnajder. GPKEX: Genetically Programmed Keyphrase Extraction from Croatian Texts. ACL 2013, pp.43, (2013)
22. J. Saratlija, J. Šnajder and B. Dalbelo-Bašić. Unsupervised topic-oriented keyphrase extraction and its application to Croatian. Text, Speech and Dialogue, pp.340–347, (2011)
23. R. Ahel, B. Dalbelo-Bašić and J. Šnajder. Automatic keyphrase extraction from Croatian newspaper articles. The Future of Information Sciences, Digital Resources and Knowledge Sharing, pp.207–218, (2009)
24. M.E.J. Newman. Networks: An Introduction. Oxford University Press.(2010)