EdgeGraph: Revisiting Statistical Measures for Language Independent Keyphrase Extraction Leveraging on Bi-grams

Amit Kumar Gupta,
Manipal University Jaipur,
Rajasthan, India.
dramitkumargupta1983@gmail.com

Muskan Garg
Mayo Clinic,
Rochester, MN, USA.
muskanphd@gmail.com

Abstract

The NLP research community resort conventional Word Co-occurrence Network (WCN) for keyphrase extraction using random walk sampling mechanism such as PageRank algorithm to identify candidate words/ phrases. We argue that the nature of WCN is a path-based network and does not follow a *core-periphery structure* as observed in web-page linking network. Thus, the language networks leveraging on bi-grams may represent better semantics for keyphrase extraction using random walk. In this work, we use bi-gram as a node and adjacent bi-grams are linked together to generate an EdgeGraph. We validate our method over four publicly available dataset to demonstrate the effectiveness of our simple yet effective language network and our extensive experiments show that random walk over EdgeGraph representation performs better than conventional WCN. We make our codes and supplementary materials available over Github[1].

1 Introduction

The language network is a textual representation of documents in the shape of a graph to exploit the best features as their characteristics. With recent developments in statistical keyphrase extraction, language network plays a pivotal role in identifying underlying patterns among words, phrases or sentences (Garg, 2021). The research community maps these patterns using the network properties as the structural properties of language networks has gained much attention in recent years (Lu et al., 2021). Existing literature contains substantial studies over the structural properties for different languages (Vera and Palma, 2021) and different domains (Garg and Kumar, 2020; Quispe et al., 2021) resulting into development of real-time language independent and domain-specific techniques, respectively.

[1] https://github.com/drmuskangarg/EdgeGraph

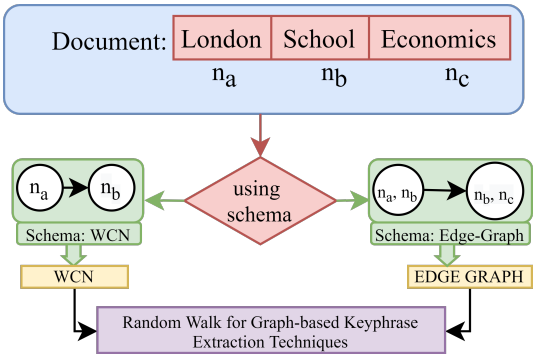

Figure 1: Overview of the proposed work

We further use the structural properties in modeling the dynamics of evolving language networks for downstream NLP applications, for instance, the Dynamic Heartbeart Graph (DHG) for event detection on Twitter (Saeed et al., 2019); and tracking the dynamics of co-word networks for emerging topics (Huang et al., 2021; Katsurai, 2017). An essential element for these graph-based topic detection and tracking applications is keyphrase extraction.

The conventional WCN is established as a benchmark representation of textual documents for random walk based keyphrase extraction(Kazemi et al., 2020; Campos et al., 2020). The random walk sampling is characterized by stochastic movement of several iterations over a network for redistributing weights to nodes. This concept of random walk was initially introduced for *web-page linking* due to the core-periphery structure (Getoor and Diehl, 2005) of the World Wide Web (WWW) connectivity. However, we observe that:

1. The WCN has significant bias towards the node which represents frequently occurring words irrespective of its context.

2. In a WCN, the edge-weight gives better insights than a node degree (Garg and Kumar, 2018a) which shows the importance of bi-gram in a language network .

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 1 - 10
December 15-18, 2022 ©2022 Association for Computational Linguistics

3. The WCN does not support the core-periphery structure like web-page linking which is an important property for the PageRank algorithm.

In this work, we study a significance of replacing WCN with EdgeGraph for random walk based GKET. The overview of our proposed approach is shown in the Fig. 1. The major contributions of this research are:

1. We propose the EdgeGraph, a graph-based textual representation to increase the information in every node and accommodate edge-distribution property.
2. We use four different publicly available text collections for keyphrase extraction to validate the EdgeGraph over WCN.
3. The statistical studies validates the effectiveness of EdgeGraph over the WCN for English dataset with medium-sized documents.

2 BACKGROUND AND RELATED WORK

The automatic keyword extraction techniques are classified into the *structured* and *unstructured* algorithms. The supervised keyword extraction is not reliable for ever-changing and dynamically evolving information (Florescu and Caragea, 2017). The unsupervised algorithms are either *statistical* or *graph-based*. A well-studied approach of supervised algorithms is graph-based keyphrase extraction.

2.1 Evolution of GKET

The PageRank algorithm (Page et al., 1999) is used for random walk sampling over *web-page linkings*. The TextRank uses PageRank algorithm and establishes itself as the first and one of the most promising random walk based GKET (Mihalcea and Tarau, 2004; Zhang et al., 2020) for textual documents. The extended version of TextRank is biased towards the node scores but explainable and is known as the Biased TextRank (Kazemi et al., 2020). The recent empirical study of TextRank (Zhang et al., 2020) shows the effectiveness of *graph-based keyphrase extraction*. PositionRank is another keyphrase extraction technique in which the position of a token plays a pivotal role (Florescu and Caragea, 2017) in identifying the candidate phrases.

The NErank (Bellaachia and Al-Dhelaan, 2012) is proposed for short-text data using *the node score* and *the edge score* over a WCN. Other than the random walk, some of the path-breaking structural GKET are *degree centrality* (Abilhoa and De Castro, 2014), *selectivity based keyword extraction* (Beliga et al., 2016), *k-core decomposition* (Tixier et al., 2016), and *keyword extraction using collective node weight* (Biswas et al., 2018) which are based on network science metrics/ models and are beyond the scope of this study. In future, the effectiveness of EdgeGraph can be studied for these structural GKET.

2.2 Historical Perspective of WCN

Graph theory has paved the path to explore language networks evolved from textual documents (Choudhury et al., 2010). The structural properties for this language network are *scale-free networks, small world property, hierarchical organization, assortativity and spectral distribution* which are studied for the WCN evolved from Chinese and English language (Liang et al., 2009), Microblogs (Garg and Kumar, 2018b), and 12 other Slavic languages (Liu and Cong, 2013). The WCN follows the *small-world property* and is *disassortative* in nature. The *eigenvalues and the spectral distribution* helps in understanding the vibrations in the linear system of language networks (Liang, 2017).

2.3 Research Gap

The semantic studies for keyword extraction techniques use Wikipedia (Wan and Xiao, 2008a), topical ranking (Awan and Beg, 2021; Bougouin et al., 2013; Boudin, 2018), and semantic connectivity (Duari and Bhatnagar, 2019). Different text-representation for semantic GKET (Osman and Barukub, 2020) are Large-scale Information Network Embedding (LINE) (Tang et al., 2015) and Context Aware Graph (CAG) (Duari and Bhatnagar, 2019). CAG incorporates the context set by two consecutive sentences by linking co-occurring words together. (Duari and Bhatnagar, 2019) use CAG for keyword extraction to eliminate the need of integer-sized sliding window parameters. Variations in weighted and unweighted adjacency matrix (Papagiannopoulou et al., 2021) and revisting this approach in literature (Ushio et al., 2021) shows that there is no existing study for variation in the text-representation with path-based network of words.

3 OUR APPROACH

In this manuscript, we propose a variation in the graph-based text representation. We find *candidate phrases* which seems to be capable to being identified as keyphrases using random walk. In this section, we discuss EdgeGraph representation.

3.1 Problem Formulation

Consider a set of pre-processed documents D as $D = d_1, d_2, d_3, \ldots d_k$ where d_i is the i^{th} document. In a document d_i, the sequence of tokens is $t_{i,1}$, $t_{i,2}$, $t_{i,3}$, $\ldots$ $t_{i,z}$, where z is the number of tokens in a document. Every token is considered as a node $t_{i,j}$ where i is the position of a document d_i and j is the position of the token in that document d_i. The total number of nodes are m and m' which varies and represents the unique number of tokens for the WCN and EdgeGraph, respectively. We use the token $(t_{i,j}, t_{i,j+1})$ as a node (n_a) in the graph for further simplification.

Definition 1: Word Co-occurrence Network (WCN): The existing WCN is a graph G of words where edges are added as (n_a, n_b). The word adjacency matrix A is created by using the co-occurrence (n_a, n_b) where the first word of the tuple (n_a, n_b) is taken as the row index and the latter word is taken as the column index in the matrix. The adjacency matrix is used to generate a WCN which is mapped as $m * m$ matrix for m: the total number of nodes in the WCN. Thus, the Graph G contains m nodes and every edge is represented as (n_a, n_b).

Definition 2: EdgeGraph: We build EdgeGraph E_G from a set of textual documents D where we map every document d_i in a graph of adjacent bi-grams. Considering a sequence n_a, n_b, n_c, the edge of a graph is the link which exists between the node (n_a, n_b) and the node (n_b, n_c) of the graph E_G. We use the bi-gram (n_a, n_b) as the node. The word adjacency matrix A' is created by using the co-occurrence $((n_a, n_b), (n_b, n_c))$ where the first node of the tuple $((n_a, n_b), (n_b, n_c))$ is taken as the row index and the latter node is taken as the column index in the matrix. The adjacency matrix is used to generate a WCN which is mapped as $m' * m'$ matrix for m': the total number of nodes in the EdgeGraph.

Given the above settings, our task is to study random walk based GKET over WCN and EdgeGraph.

Table 1: Structure of two different graph-based text representations: WCN and EdgeGraph

Graph	#Nodes	#Edges	Node: Edge	Avg Node Degree	Avg Edge Weight
WCN	581	1188	0.49	4.09	1.22
EdgeGraph	1195	1291	0.93	2.16	1.06

3.2 Problem Statement

In the WCN, the PageRank ($PR((n_b); t_q)$) of any node n_b at the time t_q depends upon the PageRank ($PR((n_a); t_{q-1})$) of the predecessor neighbours n_a of the node n_b. The idea behind this research work is to emphasise the importance of bi-gram connectivity in language network in-place of uni-grams. Thus, the PageRank, for any bi-gram $PR((n_b, n_c); t_q)$ at the time t_q depends upon the PageRank ($PR((n_a, n_b); t_{q-1})$) of the predecessor neighbours (n_a, n_b) of the node (n_b, n_c). Thus, the PageRank for E_G is represented as PR_G as shown in Equation 1.

$$PR_G(n_b, n_c) = \frac{1-d}{m'} + d \sum_{u \epsilon M(n_a, n_b))} \frac{PR_G(n_a, n_b)}{out(n_a, n_b)} \quad (1)$$

where d is the damping factor, $M(n_a, n_b)$ is a node in the set of node (bi-gram) which are directly linked to the node (n_b, n_c), and m' is total number of nodes in the EdgeGraph E_G, The EdgeGraph is evolved from $m' * m'$ adjacency matrix. The PageRank algorithm is used for random walk in the WCN (PR) and the EdgeGraph (PR_G) representation.

3.3 Working Instances

The proposed work is demonstrated over four different publicly available dataset. We use the text collection *500N-KPCrowd-v1.1* to discuss two types of working instances in this section. The first example differentiates the characteristics of the WCN and the EdgeGraph over one of the documents of *500-KP-Crowd-v1.1* dataset. The second example demonstrates the graph-based text representation of a short-text document of the dataset *500-KP-Crowd-v1.1* as WCN and EdgeGraph.

3.3.1 Example 1: Characteristics of the graph-based text representations

The nature of WCN and EdgeGraph differentiates due to uni-gram and bi-gram adjacency, respectively as shown in Table 1. The number of unique nodes (n_a) in the WCN is lesser than number of

Doc: David Mamet to debut his new play "The Anarchist" in London this year NEW YORK - A new play by Pulitzer Prize-winner David Mamet will make its debut in London this fall.

$\mathbf{d_i}$: david mamet debut new play anarchist london year new york. new play pulitzer proze winner david mamet debut london fall

Figure 2: The working instance document (Doc) and its preprocessed version d_i

Table 2: Indexing of tokens for WCN evolving from the working instance d_i

Index	Word	Index	Word
t_0	david	t_7	year
t_1	mamet	t_8	york
t_2	debut	t_9	pulitzer
t_3	new	t_{10}	prize
t_4	play	t_{11}	winner
t_5	anarchist	t_{12}	fall
t_6	london		

nodes (n_a, n_b) in the EdgeGraph because the neighbour of a word in every node may vary and unlike WCN one word may appear in more than one node in text representation of the same document. This repetition preserves the contextual difference among words with each other. The repetition of bi-grams is very limited in EdgeGraph and thus, the *node to edge ratio* is close to 1 and the *average node degree* is reduced. There is slight *increase and decrease* in *the number of edges and average edge weight*, respectively. If the number of nodes are almost doubled and there is slight increase in the number of edges; the density of the network reduces. As a result, fewer nodes with significant bi-gram are emphasized.

3.3.2 Example 2: Graph-based text representation

Consider an example of a document (Doc) which is pre-processed to obtain the document d_i as shown in Fig. 2. We index the uni-gram and bi-gram as nodes to generate the graph-based textual representation of a document d_i. The indexing of tokens are different for the WCN and the EdgeGraph as shown in Table 2 and Table 3, respectively. The WCN and EdgeGraph are generated using these indexing tables as shown in Fig. 3.

On investigating the connections of a network, we found that the important bi-gram lexical sequence is preserved in EdgeGraph and not in the WCN, for instance, *new york* and *new play* are contextually different but the word *new* is connecting both *play* and *york* in the WCN. The words *play* and *york* have different dictionary meanings and their connection does not make sense. However, in EdgeGraph, two different nodes preserve these bi-grams as the node *(new york)* and the node *(new play)*. The random walk over WCN may emphasise frequently used but unimportant words like *new* which alone does not make much sense. The EdgeGraph gives importance to meaningful bi-gram like *David Mamet*, *new play*, *new york* which make sense together. If a tuple (a,b) and (b,c) are retrieved, we combine them to form (a, b, c) and thus, n-gram keyphrases are obtained.

Table 3: Indexing of tokens for EdgeGraph evolving from the working instance d_i

Index	Word	Index	Word
t_{G0}	David Mamet	t_{G8}	New York
t_{G1}	Mamet debut	t_{G9}	play pulitzer
t_{G2}	debut new	t_{G10}	pulitzer prize
t_{G3}	new play	t_{G11}	prize winner
t_{G4}	play anarchist	t_{G12}	winner David
t_{G5}	Anarchist London	t_{G13}	debut London
t_{G6}	London year	t_{G14}	London fall
t_{G7}	year new		

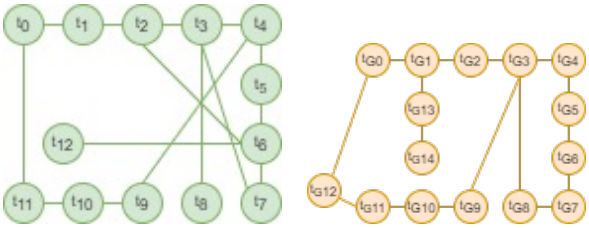

Figure 3: The WCN (left) and EdgeGraph (right) evolved from the document d_i, a working instance in Example 2

4 EXPERIMENTS AND EVALUATION

We perform the experiments with TextRank (Mihalcea and Tarau, 2004), SingleRank (Wan and Xiao, 2008b), PositionRank (Florescu and Caragea, 2017), and NErank (Bellaachia and Al-Dhelaan, 2012) over publicly available text collections. In this section, we discuss the characteristics of datasets, the experimental setup, performance evaluation and statistical significance of the proposed textual representation over the baseline.

4.1 Datasets

To test and validate the robustness of EdgeGraph over WCN, experimental results are carried out for four different datasets whose characteristics are given in Table 4. The average number of tokens per document varies from 20 to 500. The annotated

Table 4: Characteristics of the dataset used for experiments and evaluation of keyphrase extraction

Dataset	Language	Type of Doc	Domain	#Docs	#Tokens/ doc
110-PT-BN-KP	PT	News	Misc.	110	304.00
500N-KP Crowd-v1.1	EN	News	Misc.	500	408.33
pak2018	PL	Abstract	Misc.	50	97.36
wiki20	EN	Research Report	Comp. Science	20	6177.65

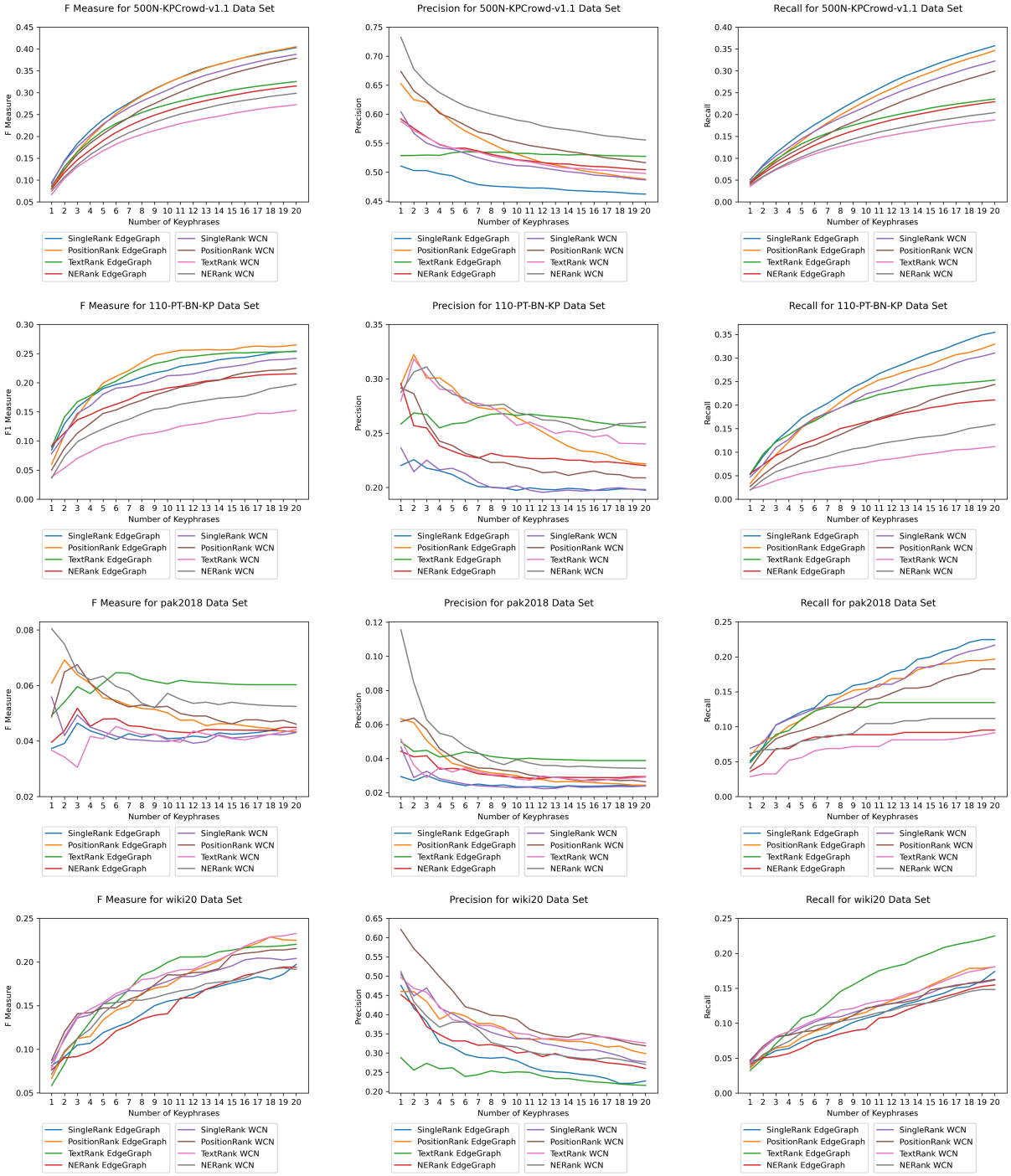

Figure 4: Performance evaluation of the random walk based GKET over WCN and EdgeGraph representation of medium-sized text for varying values of k: F-measure, Precision and Recall.

data is one of the major reasons behind variation in the resulting values of performance evaluation measures due to its subjectivity. This variation is not a potential constraint in this research work as the performance is comparative. We use four different text collections for this study: 110-PT-BN-KP (Marujo et al., 2013), 500N-KP-Crowd-v1.1 (Marujo et al., 2013), pak2018 (Campos et al., 2018), and wiki20 (Medelyan et al., 2008). Three out four dataset contains few lines of text (containing less than 500 words) to display *news* and *abstract* about miscellaneous data in three different languages. These few lines of text are different from short-text and long textual documents and thus, are termed as *medium-sized text*. The dataset *wiki2020* is in the English language which contains the *research paper* in which there are more than 4000 words in each document. We use these characteristics to categorically study the evaluation of results.

4.2 Experimental Setup

The experimental setup for this research work comprises the hardware requirements of CPU @ 2.90 GHz with Intel Core i7-7500 CPU over 64-bit Operating System having 8.00 GB RAM. We use the software of *Python 3* with library modules of *networkx* for graphical analysis, *NLTK* for text processing, *pandas* to handle the data, *matplotlib* for graph plot, and many other relevant modules.

We implement the baselines by using existing modules[2] which are further modified to incorporate the settings for EdgeGraph. We use the default parameter settings of random walk based GKET which are available in the existing implementation. The existing random walk based GKET use varying values of the sliding window parameter to generate the WCN. The most widely used value of sliding window parameter is 2 (Mihalcea and Tarau, 2004; Florescu and Caragea, 2017). The value of the damping factor d is set as 0.85 and the number of iterations are 1000. The network is converged with error rate $\epsilon < 0.01$.

As the results are comparative, we use *student's t-test* to measure the statistical significance of the results. The *Microsoft Office package* is used for the results obtained in (.csv) format to test and validate the robustness of the EdgeGraph for its statistical significance.

[2]https://github.com/boudinfl/pke

4.3 Performance Evaluation

We evaluate the performances using *Precision*, *Recall* and *F-measure* for the varying values of k where k is the number of top ranked keyphrases. All the unique tokens in extracted keyphrases are taken as the set of *extracted words*, and the tokens obtained from the ground-truth keyphrases are taken as the set of *reference words*. The performance is evaluated over these two lists: *extracted words* and *reference words* for increasing values of k over the WCN and the EdgeGraph on every dataset. The results for datasets with medium-sized text and datasets with long-text are shown in Fig. 4.

For English datasets, recall grows more steeply than non-English datasets with increasing value of k. Irrespective of language, recall shows clear improvements over EdgeGraph representation for a higher value of k as shown in Fig. 4. Precision decreases with increase in the value of k. Variation around average value of precision is lesser for medium-sized text than for long-text datasets because the probability of identifying appropriate keyphrases decreases in long-text documents due to reduced probability with large number of tokens. It is interesting to note that precision for TextRank on EdgeGraph remains constant for varying values of k.

4.4 Time Complexity

Since there is no change in the algorithm for random walk based GKET, the time complexity remains same. However, the number of nodes and the information in these nodes is increased. Also, the node to edge ratio decreases which makes the graph sparse. As the random walk is based on the Markov decision process and transition probability is increased due to change in node degree distribution.

4.5 Improvements with EdgeGraph

The experimental results are shown in Table 5. The resulting values of EdgeGraph in bold digit indicates the improvement. The datasets containing medium-sized text in which the number words are less than 500 shows better F-measure improvement over the datasets containing long text. Further, the dataset with English language shows improvement over recall in more than 90% of the cases. However, the resulting values for precision are compromised due to extraction of huge amount of data as ev-

Table 5: Results obtained for random walk based GKET over four different keyphrase extraction datasets using the WCN and the EdgeGraph text representations for $k = 20$.

Dataset 500N-KPC						
Methods	Recall		Precision		F Measure	
	WCN	EdgeGraph	WCN	EdgeGraph	WCN	EdgeGraph
Text Rank	0.1875	**0.2354**	0.4977	**0.5272**	0.2724	**0.3255**
NE Rank	0.2042	**0.2253**	0.5554	0.5042	0.2987	**0.3152**
Position Rank	0.2994	**0.3464**	0.5163	0.4878	0.3790	**0.4051**
Single Rank	0.3224	**0.3573**	0.4865	0.4621	0.3878	**0.4030**
Dataset PAK 2018						
Methods	Recall		Precision		F Measure	
	WCN	EdgeGraph	WCN	EdgeGraph	WCN	EdgeGraph 33
Text Rank	0.0913	**0.1345**	0.0291	**0.0388**	0.0441	**0.0602**
NE Rank	0.1118	0.0951	0.0342	0.0294	0.0524	0.0449
Position Rank	0.1825	**0.1966**	0.0263	0.0244	0.0461	0.0434
Single Rank	0.2166	**0.2246**	0.0239	**0.0239**	0.0430	**0.0431**
Dataset PT BN KP						
Methods	Recall		Precision		F Measure	
	WCN	EdgeGraph	WCN	EdgeGraph	WCN	EdgeGraph
Text Rank	0.2401	**0.2555**	0.2402	**0.2555**	0.2543	**0.2638**
NE Rank	0.2601	0.2201	0.2601	0.2201	0.1973	**0.2153**
Position Rank	0.2088	**0.2217**	0.2088	**0.2217**	0.2249	**0.2651**
Single Rank	0.1982	0.1974	0.1982	0.1974	0.2419	**0.2535**
Dataset WIKI 20						
Methods	Recall		Precision		F Measure	
	WCN	EdgeGraph	WCN	EdgeGraph	WCN	EdgeGraph
Text Rank	0.1809	**0.2249**	0.3258	0.2159	0.2326	0.2249
NE Rank	0.1482	**0.1547**	0.2703	0.2599	0.1914	**0.1940**
Position Rank	0.1627	**0.1805**	0.3186	0.2981	0.2154	**0.2249**
Single Rank	0.1617	**0.1740**	0.2765	0.2276	0.2041	0.1972

Table 6: Statistical Significance for different keyphrase Extraction over WCN and EdgeGraph.

Dataset 500N-KPC						
Methods	Recall		Precision		F Measure	
	t_test	p_value	t_test	p_value	t_test	p_value
Text Rank	8.8480	**3.64E-08**	-11.3725	6.38E-10	10.5934	**2.06E-09**
NE Rank	16.1771	**1.45E-12**	0.8804	***0.3896***	21.4380	**8.98E-15**
Position Rank	12.1549	**2.09E-10**	-14.8203	6.80E-12	20.5013	**2.03E-14**
Single Rank	9.2492	**1.82E-08**	-16.0827	1.61E-12	12.9572	**7.01E-11**
Dataset PAK 2018						
Methods	Recall		Precision		F Measure	
	t_test	p_value	t_test	p_value	t_test	p_value
Text Rank	4.0972	**0.00061**	-0.8953	*0.3817*	-0.4678	*0.6452*
NE Rank	23.8044	**1.31E-15**	13.5307	**3.33E-11**	27.1391	**1.17E-16**
Position Rank	-5.7231	1.62E-05	-4.0292	0.0007	-7.1178	9.08E-07
Single Rank	3.9588	**0.00084**	-1.3992	*0.1778*	-0.3888	*0.7016*
Dataset PT BN KP						
Methods	Recall		Precision		F Measure	
	t_test	p_value	t_test	p_value	t_test	p_value
Text Rank	11.3315	**6.78E-10**	-1.1464	*0.2658*	16.6051	**9.09E-13**
NE Rank	17.2039	**4.83E-13**	-0.7460	*0.4647*	30.9321	**1.02E-17**
Position Rank	22.7884	**2.93E-15**	-12.3763	1.54E-10	19.9476	**3.34E-14**
Single Rank	8.3208	**9.30E-08**	1.4777	***0.1558***	9.008	**2.75E-08**
Dataset WIKI 20						
Methods	Recall		Precision		F Measure	
	t_test	p_value	t_test	p_value	t_test	p_value
Text Rank	-5.0775	6.69E-05	-17.7666	2.71E-13	-12.0685	2.35E-10
NE Rank	5.5577	**2.32E-05**	-16.2917	1.28E-12	-1.1878	*0.2495*
Position Rank	-3.3987	0.0030	-4.0372	0.00070	-3.9784	0.0008
Single Rank	0.6545	***0.5206***	-5.6385	1.95E-05	-0.9103	*0.3740*

ery node of the EdgeGraph represents bi-gram. In this section, we analyse the results on WCN and EdgeGraph text representation for different characteristics of datasets.

4.5.1 Polish and Portuguese dataset

The random walk based GKET for Portuguese and Polish language over EdgeGraph shows major improvements with *TextRank*, *SingleRank and PositionRank*. In future, the robustness and the scalability of EdgeGraph over non-English datasets can be tested for long textual documents, different languages and for different domains.

4.5.2 Medium-sized textual documents

The English language medium-sized dataset outperforms all other datasets with EdgeGraph. It is interesting to note that though there is improvement for medium-sized textual documents, the resulting values for English and Portuguese dataset are promising but not suitable for Polish dataset.

4.5.3 Long-text documents

The long-text datasets: Wiki20, show no or slight improvement with F-measure but significant improvement with recall. The nodes represent bigrams in the EdgeGraph due to which more number of words are obtained. Hence, recall is improved more than the precision. The EdgeGraph representation gives better results over medium-sized text (containing less than 500 words) as compared to long text (containing more than 4000 words) irrespective of the language.

4.5.4 Varying number of documents

The number of documents in different dataset varies from 20 to 500 which may affect the resulting values. More the number of documents, the stronger the results. We found that the datasets with large number of documents such as 500N-KPC and 110-PT-BN-KP shows consistency over improvements for all the random walk based GKET and gives improved F-measure for all the random walk based GKET.

4.6 Statistical Significance

The results obtained by exploiting random walk based GKET over WCN and EdgeGraph are not directly comparable. We further investigate the improvements to test and validate the robustness and significance of the results. We use the *Student's t-test* with 5% of significance level. The statistical significance is evaluated over the resulting values of k varying from 1 to 20 as shown in Table 6. The null-hypothesis in $t-test$ is that the two series of resulting values are significantly different if the $p-value < 0.5$. We use the following symbolic representation for four categories of statistical analysis:

1. *EdgeGraph significantly outperforms WCN*: We represent *Bold* $p-value$ if $t-test$ is positive and the $p-value < 0.05$.

2. *EdgeGraph is better than WCN, but not statistically significant*: We represent *bold + italics* $p-value$ if $t-test$ are positive and the $p-value 0.05$.

3. *WCN is better than EdgeGraph, but not statistically significant*: We represent *italics* $p-value$ if $t-test$ are negative and the $p-value 0.05$.

4. *WCN significantly outperforms EdgeGraph*: We represent normally formatted $p-value$ if $t-test$ are negative and the $p-value < 0.05$

We investigate the improvements with similar and comparative performance of EdgeGraph over the WCN. In this context, we consider first three cases to signify non-deteriorating measure. We found that the Recall and F-measure shows good performance with EdgeGraph in 83.33% and 66.66% of the total number of cases. On the basis of individual performance, the SingleRank outperforms all other random walk based GKET.

5 Conclusion

Here in this work, we propose an EdgeGraph representation for information retrieval tasks. The experimental results over four publicly available datasets shows that keyphrase extraction is significantly improved with EdgeGraph representation leveraging on bi-grams. The recall and F- measure improves upto 27% and 18%, respectively, for the datasets with medium-sized English texts. Applicability of EdgeGraph on more than one languages (English and Portuguese) suggests its language-independencex. In future, EdgeGraph can be used for extractive text summarization, language generation and cross-lingual analysis and other information retrieval tasks. In addition to this, the massive online data can be handled using dynamics of EdgeGraph evolved from dynamically streaming data without using any pre-trained or supervised models.

References

Willyan D Abilhoa and Leandro N De Castro. 2014. A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240:308–325.

Mubashar Nazar Awan and Mirza Omer Beg. 2021. Toprank: a topicalpostionrank for extraction and classification of keyphrases in text. *Computer Speech & Language*, 65:101116.

Slobodan Beliga, Ana Meštrović, and Sanda Martinčić-Ipšić. 2016. Selectivity-based keyword extraction method. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 12(3):1–26.

Abdelghani Bellaachia and Mohammed Al-Dhelaan. 2012. Ne-rank: A novel graph-based keyphrase extraction in twitter. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 372–379. IEEE.

Saroj Kr Biswas, Monali Bordoloi, and Jacob Shreya. 2018. A graph based keyword extraction model using collective node weight. *Expert Systems with Applications*, 97:51–59.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. *arXiv preprint arXiv:1803.08721*.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. A text feature based automatic keyword extraction method for single documents. In *European conference on information retrieval*, pages 684–691. Springer.

Monojit Choudhury, Diptesh Chatterjee, and Animesh Mukherjee. 2010. Global topology of word co-occurrence networks: Beyond the two-regime power-law. In *Coling 2010: Posters*, pages 162–170.

Swagata Duari and Vasudha Bhatnagar. 2019. scake: semantic connectivity aware keyword extraction. *Information Sciences*, 477:100–117.

Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115.

Muskan Garg. 2021. A survey on different dimensions for graphical keyword extraction techniques. *Artificial Intelligence Review*, pages 1–40.

Muskan Garg and Mukesh Kumar. 2018a. Identifying influential segments from word co-occurrence networks using ahp. *Cognitive Systems Research*, 47:28–41.

Muskan Garg and Mukesh Kumar. 2018b. The structure of word co-occurrence network for microblogs. *Physica A: Statistical Mechanics and its Applications*, 512:698–720.

Muskan Garg and Mukesh Kumar. 2020. Finding summaries to obtain event phrases from streaming microblogs using word co-occurrence network. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 200–206. IEEE.

Lise Getoor and Christopher P Diehl. 2005. Link mining: a survey. *Acm Sigkdd Explorations Newsletter*, 7(2):3–12.

Lu Huang, Xiang Chen, Xingxing Ni, Jiarun Liu, Xiaoli Cao, and Changtian Wang. 2021. Tracking the dynamics of co-word networks for emerging topic identification. *Technological Forecasting and Social Change*, 170:120944.

Marie Katsurai. 2017. Bursty research topic detection from scholarly data using dynamic co-word networks: A preliminary investigation. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 115–119. IEEE.

Ashkan Kazemi, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Biased textrank: Unsupervised graph-based content extraction. *arXiv preprint arXiv:2011.01026*.

Wei Liang. 2017. Spectra of english evolving word co-occurrence networks. *Physica A: Statistical Mechanics and its Applications*, 468:802–808.

Wei Liang, Yuming Shi, K Tse Chi, Jing Liu, Yanli Wang, and Xunqiang Cui. 2009. Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and its Applications*, 388(23):4901–4909.

HaiTao Liu and Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144.

Yonghe Lu, Jiayi Luo, Ying Xiao, and Hou Zhu. 2021. Text representation model of scientific papers based on fusing multi-viewpoint information and its quality assessment. *Scientometrics*, pages 1–27.

Luis Marujo, Márcio Viveiros, and João Paulo da Silva Neto. 2013. Keyphrase cloud generation of broadcast news. *arXiv preprint arXiv:1306.4606*.

Olena Medelyan, Ian H Witten, and David Milne. 2008. Topic indexing with wikipedia. In *Proceedings of the AAAI WikiAI workshop*, volume 1, pages 19–24.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ahmed Hamza Osman and Omar Mohammed Barukub. 2020. Graph-based text representation and matching: A review of the state of the art and future challenges. *IEEE Access*, 8:87562–87583.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Eirini Papagiannopoulou, Grigorios Tsoumakas, and Apostolos Papadopoulos. 2021. Keyword extraction using unsupervised learning on the document's adjacency matrix. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 94–105.

Laura VC Quispe, Jorge AV Tohalino, and Diego R Amancio. 2021. Using virtual edges to improve the discriminability of co-occurrence text networks. *Physica A: Statistical Mechanics and its Applications*, 562:125344.

Zafar Saeed, Rabeeh Ayaz Abbasi, Muhammad Imran Razzak, and Guandong Xu. 2019. Event detection in twitter stream using weighted dynamic heartbeat graph approach [application notes]. *IEEE Computational Intelligence Magazine*, 14(3):29–38.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077.

Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. 2016. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1860–1870.

Asahi Ushio, Federico Liberatore, and Jose Camacho-Collados. 2021. Back to the basics: A quantitative analysis of statistical and graph-based term weighting schemes for keyword extraction. *arXiv preprint arXiv:2104.08028*.

Javier Vera and Wenceslao Palma. 2021. The community structure of word co-occurrence networks: Experiments with languages from the americas. *EPL (Europhysics Letters)*, 134(5):58002.

Xiaojun Wan and Jianguo Xiao. 2008a. Collabrank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969–976.

Xiaojun Wan and Jianguo Xiao. 2008b. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860.

Mingxi Zhang, Xuemin Li, Shuibo Yue, and Liuqian Yang. 2020. An empirical study of textrank for keyword extraction. *IEEE Access*, 8:178849–178858.