The CSI Journal on
Computer Science and Engineering
Vol. 17, No. 1, 2019
Pages 21-33
Regular Paper

Using Semantic Graph for Automatic Key Concept Extraction

Sudabeh Mohamadi[1] Kambiz Badie[2]

[1]Faculty of information technology, Kermanshah University of technology, Kermanshah, Iran
[2] IT research faculty, Research institute of ICT, Tehran, Iran

Abstract

Human is showered with enormous data and information. Finding the required information among them is difficult and time-consuming. Extracting Key concepts or Key phrases would assist the searcher to find the wanted information as soon as possible. In this study, a new approach is proposed regarding the Key Concept Extraction (KCE) through FrameNet lexical database. This approach is based on the natural language processing methods. The FrameNet is first applied for shallow semantic parsing of the original texts and then to construct the semantic graphs. The nodes of this graph are frames with edges of frame-to-frame relations. The concepts are weighted through the semantic graph. If the weight of concept is more than that of the threshold, it is extracted as a Key concept. The two types of the Zipfian distribution based and normal distribution based thresholds are applied here. The first outperforms the last. The human-based subjective assessment is run here.
Keywords: Semantic graph, Key concept extraction, Natural language processing, FrameNet.

1. Introduction

Due to the extraordinary growth in network and the Internet the magnitude of data in the text, image, sound and video format exceeds human being capacity. Text is one of the most abundant and practical of these categories available in books, articles, web pages, emails, organizational documents, etc.. Finding the desired data is time-consuming and sometimes impossible. In some specific applications knowing the whole data is not necessary. For example, in the search engine, the query must match the Key concepts or Key phrases of searched text; another example is finding the semantic similarity between two or more texts. If the Key concepts of one text match the Key concepts of the other, the maximum adjustment represents the most similarity between them. There exist many other applications where the Key concepts or Key phrases of text are applied. Consequently, the Key concept and Key phrase extraction are two of the most important research areas on which many methods are proposed.

Automatic Key phrase extraction is the process of identifying key terms, key phrases, key segments or keywords from a document that can appropriately represent the subject of the document [1]. The Key phrases represent the primary topics discussed in the main text, containing the words available in the main text and present a brief representation of the original text. They have the characteristic to describe and summarize the contents of documents in a compressed way [2]. Key concepts have a similar definition close to that of the Key phrases. They represent some of the underlying semantic concepts of the texts, while they consist of the words which may not always be available in the original text. The advantage of applying Key concepts rather than Key phrases is that the former are unambiguous. This leads to better understanding of key concepts than key phrases. For example, consider the following text. It is about the Dubai internet city.

Dubai 10-28 (FP) - Dubai's crown prince Sheikh Mohamed bin Rashid Al Maktoum inaugurated a free zone for e-commerce today, called Dubai internet city. The preliminary stages of the project, the only one of its kind according to its designers, are estimated at 200 million Dollar. Sheikh Mohamed, who is also the defence minister of the United Arab Emirates, announced at the inauguration ceremony that ``we want to make Dubai a new trading center.' The minister, who

has his own website, also said: ``I want Dubai to be the best place in the world for state-of-the-art technology companies. '' He said companies engaged in e-commerce would be able to set up offices, employ staff and own equipment in the open zone, including fully-owned foreign companies. The e-commerce free zone is situated in north Dubai, near the industrial free zone in Jebel Ali, the top regional and tenth international leading area in container transit. The inauguration of Dubai internet city coincides with the opening of an annual it shows in Dubai, the gulf information technology exhibition, the biggest in the middle east.

The Key concepts that can be extracted from the above text are the building internet city, the commercial competition, the e-commerce progress, the participation to foreign companies, to design of the top regional information technology zone and alike. Although they are not seen in the text, they can be understood from the context.

A new approach which can extract the Key concepts from the English texts is presented in this article. The underlying premise of Key concept extraction is interpreting the data detecting and identifying the significant concepts. The Key concept extraction is not an easy-to-do process in which even different experts may have opposing point views.

This newly proposed approach is of three major steps:

1- Interpreting the original text: since some words have different semantic role in each situation, recognizing their semantic role in the sentence is essential. The FrameNet is a powerful tool for this task, because it has a lexical database in English, based on annotating examples of how the words are used in the texts, that is human and machine-readable. More than 170,000 manually annotated sentences used in applications like information extraction, machine translation, event recognition; sentiment analysis, etc.. provide a unique training dataset for the semantic role labelling. The annotations map parts of a sentence onto the correct roles in the relevant frame. The FrameNet has annotated some continuous texts, mainly as a demonstration of how frame semantics can contribute to understanding the text [3],[4].

The situations and events have been described as semantic frames in the FrameNet database. FrameNet contains a structured network of more than 945 frames. For each frame, it also describes the relevant participants and roles. Since FrameNet was conceived as a lexical database, it offers rich linguistic information, useful for semantically describing and processing English texts. FrameNet contains more than 11,500 such pairings between a word and a frame [5].

2- To present a solution for finding the Key concepts from the framed text: to accomplish this objective, the concepts must be extracted according to the proper criteria. In this article, three scores are applied per concept: one of them is the concept frequency. The others are gained by the semantic graph, which is an undirected graph, with nodes as frames and edges as the semantic relation of the frames. The edges' weight is calculated through hierarchical distance. This graph is the sequences of semantically related frames interconnected through the semantic relations.

3- The more important concepts are extracted and returned to the main purports of the entire text: for this purpose, two types of thresholds, based on the Zipfian distribution curve and normal distribution curves are of concern. If the concept's score is more than the thresholds, it is extracted as a Key concept, otherwise, not.

To assess this approach, it is compared with the human-based Key concept extraction. The results obtained through experiments indicate that the Zipfian distribution thresholds are better than the normal distribution thresholds. Moreover, the recall and precision of all three scores are assessed.

This article is organized as follows: an overview of the related approaches is presented in Section 2; the proposed approach is explained in details in Section 3; the experimental results are presented in Section 4 and the conclusion where obtained the results are compared with the human extracted Key concepts is presented in Section 5.

2. Literature Review

Automatic Key concept extraction is one of the most important research areas in natural language processing domain. Key concepts are often applied in tasks like building ontology [6], classification [7], e-commerce [8], text summarization [9], e-content development [10], etc.. Key concepts are applied in the various domains such as mobile forensic [11], business documents [12] and medical documents [13].

The approaches of Key concept extraction are similar to that of the Key phrase extraction. Previous studies on automatic Key concept and Key phrase extraction methods are of two supervised and unsupervised categories.

Supervised methods treat the problem as a classification task, where two classes are of concern. The Key phrases are placed in class one and the remaining phrases are placed in class two. Formulation of Key phrase extraction as a supervised learning problem is introduced by Turney [14], who applied the genetic and the decision tree algorithms. He applied the two features for his algorithm: the position of word in the text and the word frequency in this formulation. The experimental results indicate that the GenEx algorithm generate better Key phrases than the C4.5 decision tree algorithm. Another notable Key phrase extraction algorithm is the KEA [15]. The authors experimented this system by applying the Naïve Bayes learning algorithm. The KEA identifies the candidate Key phrases, through the lexical methods, calculates feature values for each candidate and applies a machine-learning algorithm to predict which candidates would be good Key phrases. Kea's extraction algorithm has two stages:

1. Training: create a model for identifying Key phrases, using training documents where the author's Key phrases are known.
2.Extraction: choose Key phrases from a new document, using the above model.

The process is outlined in Figure 1. They use a large test corpus to evaluate Kea's effectiveness in terms of how many author-assigned Key phrases are correctly identified. Less than half the authors' phrases are found through KEA.

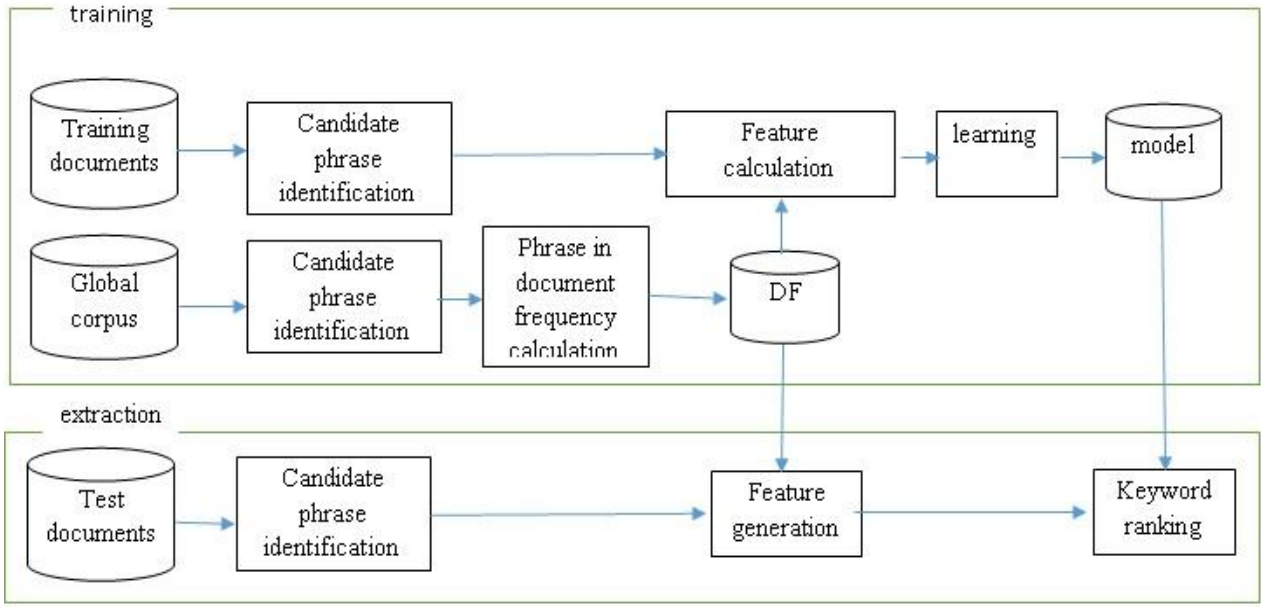

Fig. 1. The training and the extraction process in KEA [15].

A supervised methodology named SAS is proposed for Semi Automated Semantic matched concept extraction model for E-content development by Elangoven and Nirmala [10]. The researchers presented a new model for concept extraction with the objective of classification. SAS is a content based classification system, where the stemming algorithm is applied and then naive Bayes algorithm is applied as a supervised learning algorithm. By applying the probability techniques, the semantic word retrieved was used for e-content. The advantages of SAS were that very complicated grammatical rules can be applied such as the removal of multiple suffixes and prefixes. This proposed methodology is more efficient and accurate than the ones obtained through naïve based algorithms in the past.

In unsupervised methods, no data annotated with Keywords are needed; indeed, they do not require training documents where the authors provide their Key concepts or Key phrases. These methods do require data to generate information like IDF, or possibly develop a set to make some parameters or heuristic rules better. In unsupervised methods, the scoring function based on the frequency and TFxIDF is applied. The simplest unsupervised method for Key phrase extraction is the TFIDF where the candidate Key phrases are ranked according to these statistical frequencies and the top-ranked ones are selected as the Key phrases. TFIDF may miss the Key phrases with low frequencies.

One of the unsupervised systems is the CFinder [16], which is developed on ontology by identifying relevant domain concepts and their semantic correlations from a text corpus. A heuristic method that combines NLP technique, statistical knowledge, domain specific knowledge and inner structural pattern of extracted terms is applied in CFinder. CFinder consists of three steps to discover key concepts. Its overall procedural steps are outlined in Figure 2. CFinder apply the ontology named DO4MG ontology. CFinder is compared with three of the latest methods for Key concept extraction like Text2Onto, KP-Miner, and Moki, where it is found that CFinder is of better f-measure and precision value. The authors claim that real strength of CFinder lies in that it performs in a unsupervised manner which means it does not rely on training documents to build a model. Also, it does not require many corpora resources to perform compared to the corpus-based approaches that exploit multiple documents collections in multiple domains. Further, CFinder is designed to work with a corpus consisting of a small number of documents even a single document. This aspect is also an advantage of CFinder compared to the TF-IDF based approaches that typically require a large number of documents in the corpus to perform effectively.

[12] Menard and Ratte proposed an approach to extract the relevant concepts from business documents for any software project. Their system detected the Key concepts by targeting the software documentation such as manual or software requirement specification. They applied the positive and negative low-level filters' pipeline. The role of negative filters is to dismiss irrelevant or invalid expressions. The objective of this process is to achieve high precision. The positive filter modules are added to the pipeline to detect more relevant concepts; this is a recall-based process. The candidate concepts with low precision are generated in the first step, followed by reordering the concept list for higher precision. The extracted concepts are reordered through a weight propagation algorithm. When this system is tested on French corpora in public organizations, its performance is 2.7 times better than the statistical baseline relevant concept detection.

A brief survey of Key concept extraction algorithms is represented by Aman et al. [17]. They categorized the algorithms describing the necessary details and limitation of different approaches. They conducted an empirical analysis of three state-of-the-art unsupervised data driven Key concept

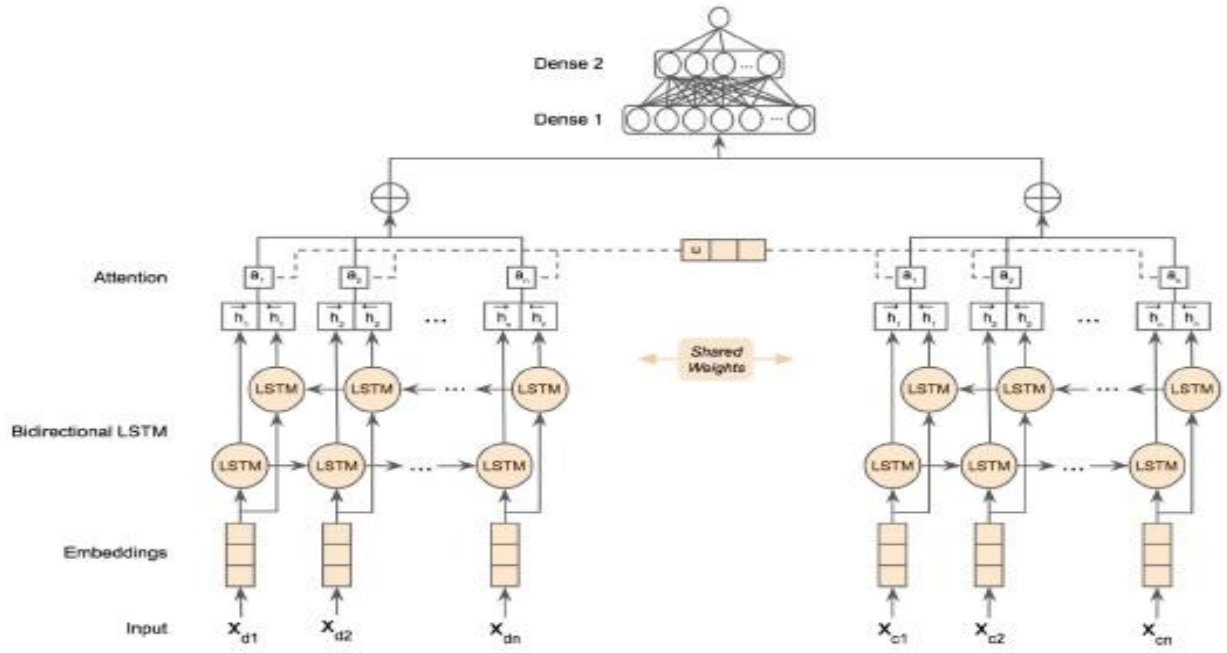

Fig. 3. The architecture of neural Key phrase classifier [20].

[25] .Mohamadi et al were the first to present an approach in extracting the Key points from English texts based on the FrameNet . The graph-based methods were adopted there , where the lexical chains were constructed through FrameNet. In lexical chains, the nodes were the frames and the edges were the frame-to-frame correlations. Then four features were computed for each one of the lexical chains which were applied in extracting significant points. These features were defined as follows:

Score 1= the number of chain member

Score 2= the sum of edges' weight

Score 3 (the hybrid score) = score 1+ score 2

Score 4= length (lexical chain) * homogeneity

$$homogeneity = 1 - \frac{number\ of\ distinct\ occurrence}{length\ (lexical\ chain)} \quad (1)$$

where , the length (lexical chain) is the number of chain members.

To extract the important chains, a threshold according to the normal distribution was applied as the Average (scores) + 2 * Standard Deviation (Scores) . The chains with the scores greater than the threshold were extracted, and their members constitute the Key points. The performance criteria were the recall and precision . This automatic system was of good recall , while the precision was low because the number of wrong extracted concepts was high. This problem was a result of extracting the lexical chains instead of concepts . The system was assessed by applying five full texts and two experts per text.

In 2017 , their attempt was made to modify the previous algorithm with a priority given to concept chains instead of lexical chains where their nodes of them were the frames [26] . The edges' weight was the semantic distance of the nodes. Then four score for each concept was calculated. Two of the scores were based on their lexical chains and the others were the concept frequency and the sum of edge's weight connected to that concept. In the final stage, three thresholds were considered for extracting significant concepts : the average (scores) + C*Standard Deviation (scores), where, C is 0, 1, and 2. The maximum precision obtained by average and the maximum recall obtained by average + 2*standard deviation . This newly proposed approach was assessed through ten full texts and five experts per texts.

3. The Proposed Approach

The architecture of the proposed approach is illustrated in a flowchart in Figure 4. The input consists of the main text, with the output as the Key concepts. The output of each stage is the input of the next. This approach is of four main stages.

3.1. Document preprocessing

The input of this stage is an original text. Its output is the XML file. This stage has two sub stages of text segmentation and semantic parsing.

3.1.1. Text segmentation

The input text must be segmented through the segmentation methods. This is the process of dividing the entire text into meaningful segments, through the segmenter tool which then breaks the text into meaningful units of words, sentences, or topics. Several applications are available for this purpose. In this study, the selected application is the Marphadorner run through two linear segmentation methods introduced by Mari Hearst's TextTiler [27] and Freddy Choi's [28].

Closeness centrality defines how close a node is to all other nodes in semantic graph and quantifies the importance of nodes based on their average distance from other nodes [32]. Closeness centrality of node C_i is the shortest path from C_i to all N-1 other nodes [33], expressed as follows:

$$closeness(C_i) = \sum_{C_i \in C \text{ and } C_j \in C} Distance\ (C_i\ .C_j) \quad (3)$$

where, $_jC$ consists of all adjacent nodes of C_i. By assuming N as the number of graph's nodes, then, the greater the value means the more central [34]. This centrality depends on |C|=N, therefore, closeness centrality is normalized by N-1.

$$Score_1(C_i) = Maximum(\frac{\sum_{j=1}^{N-1} Distance(C_i\ .C_j)}{N-1}) \quad (4)$$

The betweenness centrality is based on the idea that when a node has more participation in the shortest paths between the other nodes, it is more important in the graph. By assuming the number of shortest path between C_k and C_j is σ(C_k , C_j), then σ(C_k, C_j |C_i) is the number of shortest path between C_k and C_j passing through node C_i. The betweenness centrality of node C_i is computed as follows:

$$Score_2(C_i) = \sum_{C_k\ .\ C_j \in C}^{n} \frac{\sigma(C_k.C_j|C_i)}{\sigma(C_k.C_j)} \quad (5)$$

The semantic graph G(C , R) has not any loop and the repetitions of the concept are neglected. This metric can increase the chance of a concept to be the Key concept, therefore, the other score for each concept is considered as the concept frequency, presented as follows:

$$Score_3(C_i) = The\ frequency\ of\ C_i \quad (6)$$

An example of this is that the semantic graph has three main nodes together with three weighted edges and one intermediate node as cogitation, Figure 5. The bold line between scrutiny and research is weighted 1. The dotted-line between research and worry is weighted 1/2. The dashed-line between scrutiny and worry is weighted 1/3. The intermediate nodes do not exist in the semantic graph but they communicate relation with other nodes.

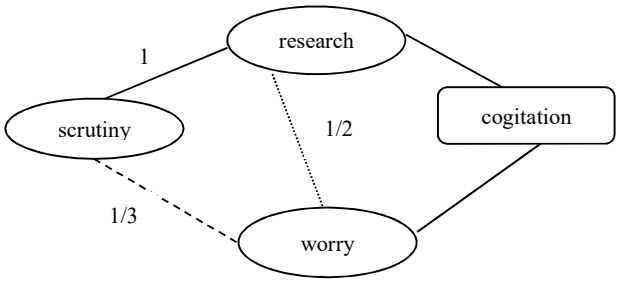

Fig. 5. The semantic graph with three main concepts.

The scores of each frame are tabulated in Table 1. The score 2 of research is 1/3, because there exist three paths on the whole and a path between worry and scrutiny passes through research node. The score 3 of each frame depends on its number of repetitions.

Table 1. Scores of the semantic graph nodes

Frame	$Score_1$	$Score_2$
scrutiny	$\frac{4}{3}$	0
research	$\frac{3}{2}$	$\frac{1}{3}$
worry	$\frac{5}{6}$	0

3.4. Key concept extraction

Here, a threshold is applied to ignore the non-important concepts and the remainders are considered as the Key concepts. In this approach, two types of thresholds according to the normal distribution and the Zipfian distribution curve are applied. If the concept score is greater than that of the threshold, then, it is a Key concept, otherwise, not.

The two thresholds of normal distribution are: the average of concept scores and the average of scores + the standard deviation of scores. According to the normal distribution curve, 50 percent of the concepts should have a score more than that of the threshold 1 and the 15.9 percent of them should have a score more than threshold 2.

The other types of thresholds are obtained through the Zipfian distribution curve. The Zipfian distribution curve refers to a distribution of probabilities of occurrences that follow the Zipf's law [35]. The Zipf's law is often adopted in linguistics, semantic and information retrieval, parser evaluation and modelling of rare events [36]. The Zipfian distribution is a discrete distribution, that is, it is empirical not theoretical, stating that the rank of words in a text is inverse in its approximate proportionality to its frequency.

Accordingly, the words of text are sorted in a descending manner with respect to their scores. Following this, numbers 1, 2, 3 … are assigned to every word as a rank. The high rank means the low frequency. In this article, the dynamic subdivision (7) is made to calculate these Zipfian thresholds as follows [24]:

$$X(i\ .n) = Score_{min} + \frac{i * (Score_{max} - Score_{min})}{2^n} \quad (7)$$

where, $_{min}$Score and $Score_{max}$ are the minimum and maximum value of the concepts' score, respectively. For every score, the three X(i, n) is calculated according to the scores' distribution.

4. Evaluation

Each one of the phrases can potentially be a Key concept, while only the ones that match human Key concept assignment are the Key concepts [14]. Accordingly, this newly proposed approach is evaluated through human extracted Key concepts. Five experts have extracted the Key concepts of each text manually. These experts are intermediate in English language and are not native. The agreement percentage is computed to assess the agreement among the experts. In this

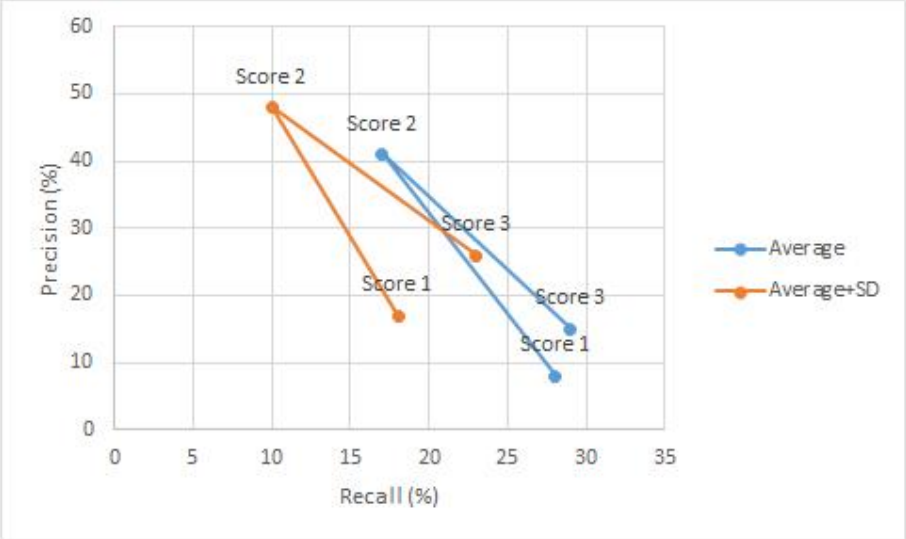

Fig. 6. Precision and Recall curve

Table4 .The recallprecision and F ,- measure of experience2 by score1.

Zipfian Thr.	Recall	Precision	F-measure
X(1,2)= 31	0.22	0.25	0.23
X(2,2)= 15.5	0.41	0.09	0.14
X(1,4)= 7.75	0.51	0.08	0.13

Table5 .The recall ,precision and F-measure of experience 2by score 2.

Zipfian Thr.	Recall	Precision	F-measure
X(1,4)= 8.81	0.18	0.55	0.37
X(1,5)= 4.9	0.22	0.37	0.27
X(3,6)= 3.9	0.23	0.31	0.26

Table6 .The recall ,precision and of experience 2by score 3.

Zipfian Thr.	Recall	Precision	F-measure
X(1,2)= 3.75	0.22	0.38	0.27
X(2,2)= 2.37	0.23	0.25	0.24
X(1,4)= 1.68	0.35	0.15	0.21

The maximum recall is 51 percent at score 1 by X(1,4) and the maximum precision is 55 percent at score 2. In general, score 1 has maximum recall and score 2 has maximum precision.

According to Tables 4 to 6, in both experiments, score 2 (betweenness centrality) has the greatest and score 1 (closeness centrality) has the lowest precision. Scores 2 and 3 obtained the close results and this phenomenon is not coincidental. As a whole, the Zipfian thresholds yield better precision, recall and F-measure than Normal Distribution thresholds, Table 6.

5. Conclusion and Future Work

In this article, an approach is proposed for the Key concept extraction. Unlike the available studies where WordNet or Wikipedia are applied, this study is the first attempt to apply FrameNet in constructing semantic graph. Each concept has three scores 1) closeness centrality, 2) betweenness centrality and 3) concept frequency. The first two scores are based on the semantic graph and the third is the number of concept occurrence in text. Two types of thresholds are discussed in this article: normal distribution based and Zipfian distribution based. As expected, the thresholds based on Zipfian Distribution provided the best results(9 see Tables 7, 8 and).

Each one of the phrases can potentially be a Key concept, while only the ones that match human Key concept assignment are the Key concepts. Accordingly, this newly proposed approach is evaluated through human extracted Key concepts. The recall and the precision are two performance criteria that are applied to evaluate this proposed approach. The most recall is obtained by closeness centrality and the best precision is obtained by betweenness centrality.

The objective of the future study would be the application of both the FrameNet and the WordNet for role labelling of the main text.

Table7 .The comparison between the recall of the Zipfian and the normal distribution thresholds.

	Score 1	Score 2	Score 3
Average of Normal thr.	0.23	0.13	0.26
Average of Zipfian thr.	0.38	0.21	0.26

Table8 .The comparison between the precison of the Zipfian and the normal distribution thresholds.

	Score 1	Score 2	Score 3
Average of Normal thr.	0.12	0.44	0.2
Average of Zipfian thr.	0.14	0.41	0.26

Table9 .The comparison between the F-measure of the Zipfian and the normal distribution thresholds.

	Score 1	Score 2	Score 3
Average of Normal thr.	0.16	0.20	0.22
Average of Zipfian thr.	0.20	0.27	0.26

References

[1] S. Belig, A. Meštrović, and S. Martinčić-Ipšić. "An overview of graph-based keyword extraction methods and approaches" *Journal of information and organizational sciences,* vol. 39. No. 1, pp. 1–20, 2015.

[2] G. Berend, "Exploiting extra-textual and linguistic information in key phrase extraction", *Natural Language Engineering,* vol. 22, no 1, pp. 73–95, 2016.

[3] J. Ruppenhofer, M. Ellswort, MR. Petruck, CR. Johnson, and J. Scheffczyk, FrameNet II: Extended theory and practice, International Computer Science Institute, Distributed with the FrameNet Data, URL http://framenet.icsi.berkeley.edu/book/book.pdf, 2006.

[4] CF. Baker, Framenet: Frame semantic annotation in practice, Handbook of Linguistic Annotation, Springer, Dordrecht, pp. 771–811, 2017.

[5] B. Lönneker-Rodman, and CF. Baker, "The framenet model and its applications", *Natural Language Engineering*, vol. 15, no. 3, pp. 415–453, 2009.

[6] S. Tonelli, M. Rospocher, E. Pianta, and L. Serafini, "Boosting collaborative ontology building with key-concept extraction", In: Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on, IEEE, pp. 316–319, 2011.

[7] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification", *Expert Systems with Applications,* vol. 57, pp. 232–247, 2016.

[8] B. Charron, Y. Hirate, D. Purcell, and M. Rezk, "Extracting semantic information for e-commerce", International Semantic Web Conference, Springer, pp. 273–290, 2016.

[9] A. Biyabangard, and MS. Abadeh, "Word concept extraction using HOSVD for automatic text summarization", AI & Robotics (IRANOPEN), IEEE, pp. 1–6, 2015.

[10] D. Elangovan, and K. Nirmala, "Semi automated semantic matched concept extraction model for e-content development", *International Journal of Applied Engineering Research,* vol. 11, no. 5, pp. 2973–2975, 2016.

[11] A. Mohammed, "Extraction of common concepts for the mobile forensics domain", Recent Trends in Information and Communication Technology: Proceedings of the 2nd International Conference of Reliable Information and Communication Technology (IRICT 2017), Springer, vol. 5, pp. 141-154, 2017.

[12] PA. Ménard, and S. Ratté, "Concept extraction from business documents for software engineering projects", *Automated Software Engineering,* vol. 23, no. 4, pp. 649– 686, 2019.

[13] M. Kholghi, L. Sitbon, G. Zuccon, and A. Nguyen, "Active learning reduces annotation time for clinical concept extraction", *International journal of medical informatics,* vol. 106, pp. 25–31, 2017.

[14] PD. Turney, "Learning algorithms for key phrase extraction", *Information retrieval,* vol. 2, no. 4, pp. 303–336, 2000.

[15] IH. Witten, GW. Paynter, E. Frank, C. Gutwin, and CG. Nevill Manning, "Kea: Practical automated key phrase extraction", Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, IGI Global, pp. 129–152, 2005.

[16] YB. Kang, PD. Haghighi, and F. Burstein, "CFinder: An intelligent key concept finder from text for ontology development", *Expert Systems with Applications,* vol. 41, no. 9, pp. 4494–4504, 2014.

[17] M. Aman, A. bin Md Said, S. Jadid Abdul Kadir, and I. Ullah, "Key concept identification: A comprehensive analysis of frequency and topical graph-based approaches", *Information*, vol. 9, no. 5, pp. 128, 2018.

[18] S. Gehrmann et al. "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives", PloS one, vol. 13, no. 2, 2018.

[19] GK. Savova et al., "Mayo clinical text analysis and knowledge extraction sys- tem (ctakes): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association,* vol. 17, no. 5, pp. 507–513, 2010.

[20] J. Villmow, M. Wrzalik, and D. Krechel, "Automatic keyphrase extraction using recurrent neural networks", International Conference on Machine Learning and Data Mining in Pattern Recognition, Springer, pp. 210–217, 2018.

[21] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge", Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, pp. 216–223, 2003.

[22] S. Gillani, and A. Kő, "Promine: a text mining solution for concept extraction and filtering", Corporate Knowledge Discovery and Organizational Learning, Springer, pp. 59–82, 2016.

[23] G. Ercan, and I. Cicekli, "Using lexical chains for keyword extraction", *Information Processing & Management*, vol. 43, no. 6, pp. 1705–1714, 2007.

[24] G. Karim, TK. Mouna, T. Lynda, and BJ. Maher, "Graph-based methods for significant concept selection", Procedia Computer Science, vol. 60, pp. 488–497, 2015.

[25] S. Mohamadi, K. Badie, and A. Moeini, "Using frame-based lexical chains for extracting key points from texts", CONTENT 2011, The Third International Conference on Creative Content Technologies, pp. 68–73, 2011.

[26] S. Mohammadi, and K. Badie, "Key concept extraction by using of FrameNet and concept chains", *Iranian Journal of Electrical and Computer Engineering,* vol. 15, pp. 64–72, 2017.

[27] MA. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages", *Computational Linguistics,* vol. 23, no. 1, pp. 33–64, 1997.
[28] FY. Choi, "Advances in domain independent linear text segmentation", Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Association for Computational Linguistics, pp. 26–33, 2000.
[29] K. Erk, and S. Pado, "Shalmaneser–a toolchain for shallow semantic parsing", Proceedings of LREC, Citeseer, Genoa, Italy, vol. 6, pp. 527-532, 2006.
[30] R. Johansson, and P. Nugues, "LTH: semantic structure extraction using non-projective dependency trees", In: Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics, pp. 227–230, 2007.
[31] D. Das et al., Semafor 1.0: A probabilistic frame-semantic parser, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2010.
[32] MK. Tarkowski et al., "Closeness centrality for networks with overlapping community structure", Thirtieth AAAI Conference on Artificial Intelligence, pp. 622-629, 2016.
[33] C. Ni, C. Sugimoto, and J. Jiang, "Degree, closeness, and betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically", In: Proceedings of ISSI, pp. 1–13, 2011.
[34] U. Brandes, SP. Borgatti, and LC. Freeman, "Maintaining the quality of closeness and betweenness centrality", *Social Networks,* vol. 44, pp. 153–159, 2016.
[35] C. Tullo, and J. Hurford, "Modelling Zipfian distributions in language", Proceedings of language evolution and computation workshop/course at ESSLLI, pp. 62–75, 2003.
[36] I. Moreno-Sánchez, F. Font-Clos, and A. Corral, "Large-scale analysis of zipf's law in English texts", PloS one, vol. 11, no. 1, 2003.
[37] O. Medelyan, IH. Witten, "Thesaurus based automatic key phrase indexing", The 6th ACM/IEEE-CS joint conference on Digital libraries, ACM, pp. 296–297, 2006.
[38] R. Barzilay, and M. Elhadad, "Using lexical chains for text summarization", Advances in automatic text summarization, pp. 111–121, 1999.

Sudabeh Mohamadi received her B.S. degree in computer engineering from Razi University, Kermanshah, Iran, in 2007, and her M.S. degrees in Tehran University, Tehran, Iran in 2011. Now she is a lecturer in Faculty of Information Technology, Kermanshah University of Technology, Kermanshah, Iran. Her research interests are natural language processing, computer networking and cloud computing.
Email: su.mohamadi@kut.ac.ir

Kambiz Badie graduated from Alborz high school in Tehran and received all his degrees from Tokyo Institute of Technology, Japan, majoring in pattern recognition. Within the past years, he has been actively involved in doing research in a variety of issues, such as machine learning, cognitive modeling, and knowledge processing & creation in general, and analogical knowledge processing, experience modeling and modeling interpretation process in particular, with emphasis on creating new ideas, techniques and contents. Out of the frameworks developed by Dr. Badie, "interpretative approach to analogical reasoning", "viewpoint oriented manipulation of concepts", "semantic fusion for phrase interpretation" and "schema satisfaction reasoning", are particularly mentionable as novel approaches to creative idea generation, which in turn have a variety of applications in developing novel scientific frameworks as well as creating potential pedagogical and research support contents. Dr. Badie is one of the active researchers in the areas of interdisciplinary and transdisciplinary studies in Iran, and has a high motivation for applying intelligent/ cognitive/phenomenological modeling methodology to the human issues. At present, he is a member of scientific board of IT Research Faculty (Full Professor) at ICT Research Institute, an adjunct professor at Faculty of Engineering Science in the University of Tehran, and in the meantime, the editor-in-chief of International Journal of Information & Communication Technology Research (IJICTR) being published periodically by ICT Research Institute and also an invited member of Iranian Academy of Science.

Email: k_badie@itrc.ac.ir

Paper Handling Data:

Submitted: 01.21.2019
Received in revised form: 11.25.2019
Accepted: 12.01.2020
Corresponding author: Dr. Kambiz Badie
ICT Research Institute

Appendix

The proposed approach is used for a sample text ("Lucorpus-v0.3 - artb_004_a1_e1_new"). Two types of thresholds are applied in two experiments. Two normal and three Zipfian distribution based thresholds are applied for each score. The key concepts are showed below their related score.

1. Inauguration *of* free ZONE$_{Locale}$ *in Dubai for* e-commerce

2. *Dubai 10-28 (FP) - Dubai 's Crown* Prince SHEIKH$_{Leadership}$ *Mohamed Bin Rashid Al Maktoum* inaugurated *a* free ZONE$_{Locale}$ *for* e-commerce TODAY$_{Calendric_unit}$, CALLED$_{Labeling}$ *Dubai Internet*CITY$_{Political_locales}$.

3. *The* PRELIMINARY$_{Preliminaries}$ stages *of the* PROJECT$_{Project}$, *the* ONLY$_{Sole_instance}$ one *of its* KIND$_{Type}$ ACCORDING$_{Attributed_information}$ TO$_{Attributed_information}$ *its* designers , *are* ESTIMATED$_{Estimating}$*at $200* MILLION$_{Cardinal_numbers}$.

4. *Sheikh Mohamed , who is also the* DEFENSE$_{Defending}$ MINISTER$_{Leadership}$ *of the United Arab Emirates ,* ANNOUNCED$_{Statement}$ AT$_{Locative_relation}$ *the* inauguration ceremony *that `` we* WANT$_{Desiring}$ *to*MAKE$_{Cause_change}$ *Dubai a* new trading CENTER$_{Locale_by_use}$. ''

5. *The* MINISTER$_{Leadership}$, *who* HAS$_{Possession}$ *his own website , also* SAID$_{Statement}$ *: `` I* WANT$_{Desiring}$ *Dubai to be the* BEST$_{Usefulness}$ PLACE$_{Locale}$ IN$_{Interior_profile_relation}$ *the* WORLD$_{Political_locales}$ *for*state-of-the-art TECHNOLOGY$_{Gizmo}$ COMPANIES$_{Businesses}$. ''

6. *He* SAID$_{Statement}$ COMPANIES$_{Businesses}$ ENGAGED$_{Intentionally_act}$ *in* e-commerce *would be* ABLE$_{Capability}$ *to* SET$_{Intentionally_create}$ UP$_{Intentionally_create}$ OFFICES$_{Building_subparts}$, EMPLOY$_{Employing}$ staff *and* OWN$_{Possession}$ equipment *in the* OPEN$_{Openness}$ zone , INCLUDING$_{Inclusion}$ fully-owned FOREIGN$_{Foreign_or_domestic_country}$ COMPANIES$_{Businesses}$.

7. *The* e-commerce free ZONE$_{Locale}$ *is* SITUATED$_{Being_located}$ IN$_{Interior_profile_relation}$ NORTH$_{Part_orientational}$ *Dubai ,* NEAR$_{Locative_relation}$ *the* INDUSTRIAL$_{Manufacturing}$ free ZONE$_{Locale}$IN$_{Interior_profile_relation}$ Jebel *Ali , the* top REGIONAL$_{Locale}$ *and* tenth INTERNATIONAL$_{Foreign_or_domestic_country}$ LEADING$_{First_rank}$ AREA$_{Locale}$ *in* CONTAINER$_{Containers}$ transit .

8. *The* inauguration *of Dubai Internet City* COINCIDES$_{Simultaneity}$ *with the* opening *of an* ANNUAL$_{Frequency}$ *IT* show IN$_{Interior_profile_relation}$ *Dubai , the Gulf Information Technology* Exhibition *(Gitex) , the*biggest *in the Middle East* .

Normanl distribution based

Threshold 1= Average(Scores)			**Threshold 2= Average(Scores)+SD(scores)**		
Score 1	Score 2	Score 3	Score 1	Score 2	Score 3
Being_located Building_subparts Calendric_unit Capability Containers Desiring First_rank Frequency Intentionally_act Intentionally_create Locale Locale_by_use Locative_relation Political_locales Possession Statement Usefulness	Frequency Intentionally_create Locale Political_locales	Businesses Desiring Foreign_or_domestic_country Interior_profile_relation Leadership Locale Locative_relation Political_locales Possession Statement	Being_located Frequency Intentionally_act Intentionally_create Locale_by_use Locative_relation	locale	Businesses Interior_profile_relation Leadership Locale Statement

Zipfian distribution based

Score1			Score2			Score3		
X(1,2)	**X(2,2)**	**X(1,4)**	**X(1,4)**	**X(1,5)**	**X(3,6)**	**X(1,2)**	**X(1,3)**	**X(1,4)**
Frequency Intentionally_act Locale_by_use	Being_located Building_subparts Calendric_unit Capability Containers	Attributed_information Being_located Building_subparts Calendric_unit	Locale	Locale Political_locales	Frequency Intentionally_create Locale Political_locales	Interior_profile_relation Locale	Businesses Interior_profile_relation Leadership Locale Statement	Businesses Desiring Foreign_or_domestic_country Interior_profile_relation Leadership Locale

	Desiring First_rank Frequency Intentionally_act Intentionally_create Locale Locale_by_use Locative_relation Political_locales Possession Statement Usefulness	Capability Cause_change Containers Desiring Estimating First_rank Foreign_or_domestic_country Frequency Gizmo Intentionally_act Intentionally_create Interior_profile_relation Leadership Locale Locale_by_use Locative_relation Manufacturing Openness Political_locales Possession Project Statement Usefulness						Locative_relation Political_locales Possession Statement