






Applying OWA Operator in the Semantic Processing for Automatic Keyphrase Extraction

Manuel Barreiro-Guerrero¹, Alfredo Simón-Cuevas¹ ,
Yamel Pérez-Guadarrama², Francisco P. Romero³ ,
and José A. Olivas³ 

¹ Universidad Tecnológica de La Habana José Antonio Echeverría,
Ave. 114, No. 11901, Marianao, La Habana, Cuba
{mbarreiro, asimon}@ceis.cujae.edu.cu

² Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), 7ma A,
No. 21406, e/214 y 216, Playa, La Habana, Cuba
yperez@cenatav.co.cu

³ Universidad de Castilla-La Mancha, Paseo de la Universidad, 4,
Ciudad Real, Spain
{FranciscoP.Romero, JoseAngel.Olivas}@uclm.es

Abstract. The automatic keyphrases extraction from texts is a useful task for many computational systems in the natural language processing and text mining fields. Although several solutions to this problem have been developed, the semantic analysis has been one of the linguistic features less exploited in the most reported proposal, causing that the obtained results still show low accuracy and performance rates. This paper presents an unsupervised method for keyphrase extraction, which is based on the use of lexical-syntactic patterns for extracting information from texts and a fuzzy modelling of topics. An OWA operator which combines several semantics measures has been applied in the topic modelling process. This new approach was evaluated with Inspec and 500N-KPCrowd datasets and compared with other reported systems, obtaining promising results.

Keywords: Automatic keyphrase extraction · Linguistic patterns · Topic modelling · Semantic processing · OWA operator

1 Introduction

The exponential growth of textual and unstructured data in digital format have provoked that to distill the most relevant information from the amount of available information constituted a significant challenge to the textual information processing. The development of computational solutions based on the application of natural language processing (NLP) and text mining techniques emerging as the most promising alternatives to deal with this challenge. In this context, a high-level description of a document can be achieved through relevant words or phrases, by its strong relationship with the main topic (s) that are addressed in the documents, so that the automatic

keyphrase extraction constitute an essential task for many text mining solutions [3, 10]. The keyphrase provides a concise understanding of a text, enabling one to grasp the central idea and the main topics discussed in a text document and facilitates the construction of text representation models, as graph-based models.

Several automatic keyphrases extraction solutions have emerged over the last few years, some following a supervised approach [8, 9] and others following unsupervised techniques [1, 6, 11, 14–17]. In this work we focused on unsupervised keyphrase extraction, where a human-annotated training data for applying some machine learning algorithm do not require in this process. The solutions reported still show low rates of accuracy and performance [3, 10], and the semantic constitutes one of the linguistics feature less exploited in the most reported proposal, fundamentally in unsupervised approaches. According to [10], it is essential to focus on semantically and syntactically correct phrase aspects and make sure that the keyphrases are semantically relevant to the document topic and context. Topic modelling for keyphrase extraction from texts has been reported in [1, 14, 15], however the semantic analysis in those proposals has not been considered or not in all its possible dimensions, constituting a weakness. The semantic analysis of textual content, at the level of words meaning or relationships among them, is usually subject to subjectivity, vagueness and imprecision problems, due to the inherent ambiguity of the natural language, which constitutes a challenge for the computational solutions that requires intensive semantic processing. The fuzzy logic offers several techniques for dealing with these problems, such as fuzzy set techniques, fuzzy clustering algorithms, aggregation operators, and others. Despite these advantages, few keyphrase extraction proposals that apply a fuzzy logic approach to carried out some semantic analysis level have been identified [15].

In this paper, an unsupervised method for automatic keyphrase extraction from a single document is proposed. The method was conceived through the combination of the use of lexical-syntactic patterns with a graph-based topic modeling, which is carried out from the fuzzy logic perspective. In this sense, syntactic and semantic measures are combined, applying the aggregation operator OWA (Ordered Weighted Averaging) [18], to increase the semantic processing level of the candidate phrase in the topic identification. The method was evaluated with the Inspec [8] and 500N-KPCrowd [9] datasets, and the performance was measured using the precision, recall, and F-score metrics. The results were compared with those obtained by other state-of-the-art unsupervised proposals, reaching improvement the results respect to those systems included in the comparison. Concretely, the contributions of this paper are the following: (1) we propose a new way for processing the semantic information in topic modelling based keyphrase extraction solutions, applying a fuzzy aggregation operator (OWA), and (2) we prove on two datasets that the fuzzy topic modelling proposed can improve the accuracy in the unsupervised automatic keyphrase extraction process.

The rest of the paper is organized as follows: Sect. 2 summarizes the analysis of the related works; Sect. 3 describes the proposed method; Sect. 4 presents the experimental results and the corresponding analysis. Conclusions and future works are given in Sect. 5.

2 Related Works

The solutions of automatic keyphrase extraction in text documents are usually designed in 4 phases: pre-processing, identification and selection of candidate phrases, keyphrase determination, and evaluation [10]. The unsupervised approach has the advantage of using only the contained information in the input text to determine the keyphrases, and several solutions have been reported [1, 6, 11, 14–17].

In TopicRank [1], a strategy based on the identification and analysis of topics to extract the relevant phrases is proposed. In this method, the longest sequences of nouns and adjectives in the text are extracted as candidate phrases, and the syntactically similar noun phrases are clustered into a theme or topic, using a hierarchical agglomerative clustering (HAC) algorithm [12]. Next, a graph is constructed where each vertex represents a topic and the arcs are labeled with a weight that represents the strength of the contextual relationship in the text among the contained candidate phrase in a topic regarding those that were grouped in another topic to which it relates. Finally, the selection of only one keyphrase from each topic is carried out, which is a weakness because a topic can be represented by more than one keyphrase in the same text. This proposal is improved in [14], through conceiving a more flexible procedure of keyphrase selection from topics and incorporating the definition of a distance between phrases function in the candidate-phrases clustering process, although semantic processing remains limited, as in the case of [1]. Liu et al. [6] also consider the clustering of candidate phrases to represent the document themes, and a cooccurrence-based relatedness measure is applied for computing the semantic relatedness of candidate terms in this process.

In TextRank [11] the candidate terms and their relationships are represented in an unweighted and undirected graph, whose vertexes represent the terms and the arcs represent co-occurrence relationships between them. An algorithm similar to PageRank [2] is applied to the constructed graph for determining the relevance of each vertex. Next, the third part of the vertexes of the whole graph is selected as the most relevant vertexes. Finally, the relevant terms are marked in the text and the sequences of adjacent words are selected as keyphrases. A similar solution is considered in the Saliency Rank algorithm [17], but the use of PageRank [2] to obtain a ranking of the words in the document is combined with other word saliency measures in the context of an LDA (Latent Dirichlet Allocation) based topic modelling approach. In [15], the co-occurrence graph of the input text's words is created, which is customized for each topic by employing the semantic information obtained from the topic model (built over the Wikipedia's articles) to form the topical graphs. Next, the communities and central nodes of these topical graphs are detected. In this process, the fuzzy modularity criterion for measuring the goodness of overlapped community structures is applied. The co-occurrence graph is also applied in RAKE [16]. In this approach, the graph is constructed with all individual words founded in the candidate keyphrases, and used to calculate the scores of each word and keyphrase. The word score is calculated through the word degree as well as the word frequency. For multiple-word expressions, they calculated the weights by summing the members' weights up.

According to the related works analyzed, the graph-based terms representation and the topic modelling appear as promising alternatives for the unsupervised keyphrase extraction from text. The unsupervised methods offer more significant strengths than supervised ones; nevertheless, have as a weakness that graph-based approach not guarantee that the extracted keyphrases represent all the main topics of the document and fail to reach a reasonable coverage level of the text document [10]. The good keyphrases of a document should be semantically relevant to the document theme or topic and cover the whole document well [6]. In this sense, can be seen in the analyzed works a low use of the semantic analysis in the clustering and topic modelling process carried out or in other task included. This semantic processing has focused on computing the co-occurrence relatedness [6, 11, 15, 16] or distance-based contextual relationship [14]. However, there are other semantic analysis level and measures, such as: semantic similarity and semantic relatedness measures, which have not been explored. Our work is aimed at assessing the benefits of these other semantic measures in the topic modelling from the fuzzy logic perspective to improve the outcomes of the unsupervised keyphrases extraction process.

3 Keyphrase Extraction Using OWA Operator

The proposed method was conceived through the combination of the use of lexical-syntactic patterns with a topic modelling carried out from a fuzzy perspective. This method has four phases: (1) text pre-processing, (2) fuzzy identification of topics, (3) relevance evaluation of topics, and (4) keyphrases selection. The lexical-syntactic patterns were defined for extracting candidate phrases from the text, and a fuzzy clustering of candidate phrases is proposed for identifying the main topics in the texts, to increase the semantic analysis in this process respect other proposals [1, 14, 15]. Also, a more flexible mechanism of keyphrases selection from the relevant topics identified is incorporated, that allows extracting more than one keyphrase and solving the weakness identified in TopicRank [1].

3.1 Text Pre-processing

In this phase, different NLP tasks are carried out for extracting the syntactic information from the text, which is required in the candidate phrases extraction process. Initially, plaintext from the input file is extracted, segmented into paragraphs and sentences, and the set of tokens (e.g., words, numbers, and others) are obtained from each sentence. Subsequently, the deep syntactic analysis using the Freeling parser is performed. The extraction of candidate phrases is based on the identification of conceptual phrases and a set of defined lexical-syntactic patterns are defined for this purpose, such as: [D | P | Z] + [<s-adj>] + NN; [D | P | Z] + [<s-adj>] + NN + NN; [Z] + <sn>; NN + [IN] + NN; JJ + NN + [NN], in a similar way to that reported in [14]. These patterns have been defined according to the grammar labeling used by Freeling, and they combine a set of relevant grammatical categories in the composition of concepts. Most of these patterns have their origins in the most frequent patterns identified in the concepts included in several ontological knowledge resources analyzed, e.g. the ontology of the

DBpedia project [4], which has more than 1000 concepts of different domains from Wikipedia. Through these patterns the coverage of the text in this process is increased, respect to other proposals that only consider noun phrases.

3.2 Fuzzy Identification of Topics

The topics identification process is carried out using a hierarchical agglomerative clustering algorithm [12] of the extracted candidate phrases, which is addressed as a fuzzy logic problem for reinforcing the semantic analyses in the phrases clustering. Although the use of clustering algorithms for topics modelling has also been reported in [1, 14, 15], the semantic analysis in those proposal has not been considered or not in all its possible dimensions. This is a weakness considering the assumption that a topic could be modelled through the cluttering of concepts that frequently appear together as well as concepts with similar meanings or semantically related. To address this weakness, in our new unsupervised approach the phrases clustering process is carried out considering the resultant score of combining the syntactic similarity and distance between phrases measures reported in [14] with other two semantic similarity measures applying a fuzzy aggregation operator. These two semantic similarity measures were conceived according to the sentence-to-sentence similarity metric reported in [5] and using two word-to-word semantic similarity-relatedness metrics from WordNet::Similarity package, specifically the Jiang & Conrath and Leacock & Chodorow metrics [13]. Additionally, the words distance metric reported in [14] was redefined (Eq. 1):

$$D(F_1, F_2) = \begin{cases} 1 & \text{if } F_1 \text{ and } F_2 \text{ appear in the same paragraph} \\ 1 - \frac{\text{ave_dist}(F_1, F_2)}{TW} & \text{in other cases} \end{cases} \quad (1)$$

where $\text{ave_dist}(F_1, F_2)$ is the average distance [14] in words that exists between the words included in the pair of phrases F_1 and F_2 , and TW is the total of words in the text.

In this method, the OWA operator [18] is applied for aggregating the resultant numerical values (a_i) from the four defined measures into a single one similarity-relatedness score (SRS) of a pair of candidate phrases. These measures represent features with different semantic “meaning” for the phrases clustering, as well as different relevance levels for the decision making in this process. OWA operators are very useful for dealing with such problems, modelling the semantics and relevance levels through weights assigned to each measure. To combine these syntactic, distance and semantic measures using an OWA Operator allows to achieve clusters of phrases strongly related among them from different semantic dimensions, and at the same time, to achieve a wide coverage of the whole document in the topic modelling process.

Definition: An OWA operator of dimension n is a mapping denoted $f_{owa} : \mathbb{R}^n \rightarrow \mathbb{R}$ that has associated an n -dimensional weight vector $W = [w_1, w_2, \dots, w_n]^T$ such as $w_i \in [0, 1]$ and $\sum_i w_i = 1$. The function f_{owa} is defined according to Eq. 2, with b_j the j th largest element in the collection $a_1 \dots a_n$.

$$f_{owa}(a_i, \dots, a_n) = \sum_{j=1}^n w_j b_j \quad (2)$$

There are different methods for determining the weights to be used in an OWA operator and the use of linguistic quantifiers is one of them [20], e.g. RIM (Regular Increasing Monotone) quantifiers. Yager proposed a method to calculate the weights of an operator OWA by means of RIM quantifiers [19], which is defined in Eq. 3. Specifically, in our proposal, we apply the RIM quantifier “Most (Feng & Dillon)” reported in [7] (see Eq. 4), as the first approach to measure the performance of the OWA operator in the keyphrase extraction problem.

$$w_j = Q\left(\frac{j}{n}\right) - Q\left(\frac{j-1}{n}\right) \quad (3)$$

$$Q(x) = \begin{cases} 0 & \text{si } 0 \leq x \leq 0.5 \\ (2x - 1)^{0.5} & \text{si } 0.5 < x \leq 1 \end{cases} \quad (4)$$

The hierarchical agglomerative clustering process is carried out by means of creating a square symmetric matrix of size n (total of candidate phrases identified), where each topic identifies a row, and a column and the intersection between each pair of topics contains the SRS (weight value) between a pair of candidate phrases that represent the corresponding topics. Initially, each candidate phrase is considered as a topic. In each iteration, the pair of topics with the highest weight value is clustered. The weight values average is used as a clustering strategy of a pair of topics, according to the reported in TopicRank [1]. The phase concludes generating a graph representation of the text, in which the identified topics are represented as vertices and these are linked by labeled arcs with the weight of the relation between them. Each weight represents the strength of the existing semantic relationship between the pair of topics. The topics A and B have a strong semantic relationship if the candidate phrases that includes those topics which frequently appear closer in the text. The weight W_{ij} is calculated according to Eqs. (5) and (6). Equation (6) refers to the reciprocal distance between the positions of the candidate phrases c_i and c_j in the text, where $pos(c_i)$ represents all positions (p_i) of c_i .

$$W_{i,j} = \sum_{c_i \in T_i} \sum_{c_j \in T_j} D(c_i, c_j) \quad (5)$$

$$D(c_i, c_j) = \sum_{p_i \in pos(c_i)} \sum_{p_j \in pos(c_j)} \frac{1}{|p_i - p_j|} \quad (6)$$

3.3 Relevance Evaluation of Topics

In this phase, the relevance of each topic represented in the constructed topics graph is evaluated using the TextRank [11] model. The relevance score computed to each topic T_i is based on the concept of “voting” (inspired in the PageRank algorithm [2]): the adjacent topics of T_i with the highest score contribute more to the relevance evaluation

of the topic T_i . The relevance score $S(T_i)$ is obtained through the Eq. 7, where V_i is the set of adjacent topics of T_i in the graph, and λ is a muffled factor that usually is 0,85 [2].

$$S(T_i) = (1 - \lambda) + \lambda * \sum_{T_j \in V_i} \frac{W_{ij} * S(T_j)}{\sum_{T_k \in V_j} W_{j,k}} \quad (7)$$

3.4 Keyphrases Selection

The selection of keyphrases from the most relevant topics identified in the previous phases is carried out according to the following criteria: (1) candidate phrase that first appears in the text; (2) most frequently used candidate phrase; and (3) candidate phrase that has more relationship with the others of each topic (centroid role). A mechanism that allows combining the three criteria has been implemented in our proposal, offering the possibility of extracting more than one keyphrase from each topic and greater flexibility in its execution, respect to the reported in [1] (only one of the criteria is considered affecting the coverage in the keyphrases extraction process). If more than one candidate phrase (associated with a topic) with the same higher frequency is identified, and the frequency value is higher than 1, then all of them are selected. Otherwise, only the first candidate phrase that appears in the text will be chosen.

4 Experimental Results

The proposed method was evaluated using the Inspec [8] and 500N-KPCrowd [9] datasets, which contain texts collections written in English. Inspec is a collection of 500 paper abstracts of Computer Science & Information Technology journals with manually assigned keyphrases by the authors. 500N-KPCrowd contains 450 broadcast news stories from 10 different categories and is considered to see how the proposal perform on texts of general domain. The performance of the method was measured using the precision (P), recall (R), and F-score (F) metrics and the obtained results were compared with those obtained by others unsupervised methods reported, which have been evaluated with the selected datasets.

As shown in Table 1, the proposed method reached higher values in most metrics and both datasets, respect to those obtained by the state-of-the-art proposals compared. The best results were obtained with Inspec, where the method achieved a better balance between precision and recall, which is a very challenging target to reach and convenient for the increasing the applicability of this type of solutions. In this case, it's evidenced that the proposed fuzzy approach for the semantic processing and topic modelling not only contributed to increase the accuracy in the keyphrase extraction, but also increase the recall, which obtained result was very encouraging (near to the 60%).

The achieved recall with 500N-KPCrowd was the less satisfactory results of our proposal, although the results of precision and F-score were significantly better than those obtained by the other proposals. Although the recall of TSAKE [15] is the highest in the case of 500N-KPCrowd, its precision is approximately 30% lower than our method, and in the same way the F-score (9% lower).

Table 1. Experimental results with Inspec and 500N-KPCrowd datasets

Systems	Inspec			500N-KPCrowd		
	P	R	F	P	R	F
TextRank [11]	31.2	43.1	36.2	26.5	6.3	10.3
TopicRank [1]	36.4	39.0	35.6	26.2	23.9	25.0
TSAKE [15]	40.1	20.3	26.9	14.3	46.6	21.9
Saliency Rank [17]	26.5	29.8	26.6	25.3	22.2	22.9
RAKE [16]	33.7	41.5	37.2	12.0	3.8	5.8
Method proposed	42.1	59.9	47.9	45.5	22.8	30.8

The low value obtained of recall can be derived from the presence of a high number of annotated named entities as keyphrase in 500N-KPCrowd. The identification of named entities as candidate phrase from the text was not considered within the defined patterns in the pre-processing phase of the proposed approach, because this type of sentence is not identified often as a keyphrase. On the other hand, the OWA operator applied in the proposed fuzzy modelling of topics includes the aggregation of several semantic measures, which may fail in the case of named entities. This situation suggests specific analysis for this type of phrases in the next approaches of our proposal. Nevertheless, through the experiments carried out, the achieved effectiveness improvement by our method and the fuzzy-based semantic processing proposed in the automatic keyphrase extraction from two types of texts, such as: paper abstracts and news stories, has been demonstrated.

5 Conclusions and Future Works

In this paper, a new unsupervised method for automatic keyphrase extraction from text was presented, in which the use of lexical-syntactic patterns to identify the candidate phrases was combined with a fuzzy modelling of topics. The use of the linguistic patterns allowed to increase the possibilities for identifying the candidate phrases, and the coverage of the text. Several syntactic and semantic measures to modeling the most relevant linguistics features of the candidate phrase were aggregated applying an aggregation operator OWA. The aggregation of these measures through the OWA operator allowed to increase the semantic processing of the candidate phrase in the topic identification, which is a little-considered aspect in most of the reported proposals. The proposed method was evaluated on two datasets with different types of texts, and the obtained results were compared with those obtained by other unsupervised schemes. The most significant results were obtained on Inspec, where a better balance between precision and recall was achieved, at the same time that their values were higher than the obtained by other proposals. These metrics were also improved on 500N-KPCrowd, although the recall must be enhanced. Considering the obtained results, the proposed method reached higher values in most of the metrics, demonstrating the contribution of the applied fuzzy topic modeling for improving the

keyphrases extraction process, in paper abstract and in more general domain texts, such as the news stories.

The improvement of the recall results on general domain texts will be one of the challenges to solve in the future, considering specific analysis for the named entities. Additionally, others linguistic quantifiers applied to the OWA operator will be evaluated for measuring their performances in the keyphrase extraction process.

Acknowledgments. This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET: TIN2016-76843-C4-2-R (AEI/FEDER, UE).

References

1. Bougouin, A., Boudin, F., Daille, B.: TopicRank: graph-based topic ranking for keyphrase extraction. In: Proceedings of the 6th International Joint Conference on NLP, pp. 543–551 (2013)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), 107–117 (1998)
3. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: a survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the ACL, pp. 1262–1273 (2014)
4. Lehmann, J., et al.: DBpedia - a large-scale multilingual knowledge base extracted from Wikipedia. *Semant. Web J.* **1**, 1–27 (2012)
5. Li, Y., McLean, D., Bandar, Z.A., O’Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.* **18**, 1138–1150 (2006)
6. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to find exemplar terms for keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in NLP, pp. 257–266 (2009)
7. Liu, X., Han, S.: Orness and parameterized RIM quantifier aggregation with OWA operators: a summary. *Int. J. Approx. Reason.* **48**(1), 77–97 (2008)
8. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in NLP, pp. 216–223 (2003)
9. Marujo, L., Ribeiro, R., de Matos, D.M., Neto, J.P., Gershman, A., Carbonell, J.: Key phrase extraction of lightly filtered broadcast news. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS (LNAI), vol. 7499, pp. 290–297. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32790-2_35
10. Merrouni, Z.A., Frikh, B., Ouhbi, B.: Automatic keyphrase extraction: an overview of the state of the art. In: Proceedings of the 4th IEEE International Colloquium on Information Science and Technology, pp. 306–313 (2016)
11. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in NLP, pp. 404–411 (2004)
12. Müllner, D.: Modern hierarchical, agglomerative clustering algorithms. *CoRR*, abs/1109.2378 (2011)
13. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity - measuring the relatedness of concepts. In: Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004), pp. 1024–1025 (2004)
14. Pérez-Guadarrama, Y., Rodríguez, A., Simón-Cuevas, A., Hojas-Mazo, W., Olivas, J.A.: Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos. *Procesamiento del Lenguaje Natural* **59**, 39–46 (2017)

15. Rafiei-Asl, J., Nickabadi, A.: TSAKE: a topical and structural automatic keyphrase extractor. *Appl. Soft Comput. J.* **58**, 620–630 (2017)
16. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, pp. 1–20 (2010)
17. Teneva, N., Cheng, W.: Saliency rank: efficient keyphrase extraction with topic modeling. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 530–535 (2017)
18. Yager, R.R.: On ordered weighted averaging operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.* **18**, 183–190 (1988)
19. Yager, R.: Quantifier guided aggregation using OWA operators. *Int. J. Intell. Syst.* **11**, 49–73 (1996)
20. Zadeh, L.A.: A computational approach to fuzzy quantifiers in natural languages. *Comput. Maths. Appl.* **9**, 149–184 (1983)