Multi-document Arabic Text Summarization based on Thematic Annotation

Amina Merniz, Anja Habacha Chaibi and Henda Hajjami Ben Ghézala
National School of Computer Science, University of Manouba, Tunisia

Keywords: Text Summarization, Multi-document Summarization, Pagerank Algorithm, Thematic Annotation.

Abstract: Reduce document(s) by keeping keys and significant sentences from a set of data is called text summarization. It has been around for a long time in natural language processing research, it is improving over the years due to a considerable number of methods and research in this area. The paper suggests Arabic multi-document text summarization. The originality of the approach is that the summary based on thematic annotation such as input documents are analyzed and segmented using LDA. Then segments of each topic are represented by a separate graph because of the redundancy problem in multi-document summarization. In the last step, the proposed approach applies a modified pagerank algorithm that utilizes cosine similarity measure as a weight between edges. Vertices that have high scores are essential. Therefore, they construct the final summary. To evaluate summary systems, researchers develop serval metrics divided into three categories, namely: automatic, semi-automatic and manual. This study research chooses automatic evaluation methods for text summarization, mainly Rouge measure (Rouge-1, Rouge-2, Rouge-L, and Rouge-SU4).

1 INTRODUCTION

The quantity of data on the internet is growing continuously, which increases the appearance of redundant and unnecessary documents (data). At the time, a user is searching for accurate and relevant information (news, papers). Automatic document text summarization has become one of the active tasks in this field. It is considered a real challenge in NLP (Natural Language Processing) (even in Text mining). It serves in several scopes such as social media marketing, newsletters, email overload, and medical cases. It allows the production of a short version from the input text(s), containing the main ideas and the relevant information. Automatic document text summarization can be abstractive or extractive. The abstractive summary involves a detailed analysis of the document. It may include a new sentence not present in the initial text. However, the extractive one depends on the identification and extraction of the most frequent units of the source text to be integrated into the generated summary. A summary can also be a multi-document or mono-document, the multi-document generates the summary from the collection of documents, whereas a mono-document summarizes one document in the input. Concerning the language, text summarization can be classified into multilingual, monolingual or cross-lingual. Despite the researcher's efforts and the considered number of approaches that have been developed until our days in this field, multi-document text summarization studies in the Arabic language are still limited. (Conroy et al., 2013) produced text summarization in Arabic using latent semantic analysis method (LSA) as well as Latent Dirichlet Allocation (LDA), also (Azmi and Al-Thanyyan, 2012) proposed an Arabic summary for single document based on Lexical cohesion, (El-Haj and Rayson, 2013) introduced statistical measures to produce mono and multi-document text summarization in both Arabic and English languages, however, (Fejer and N.Omar, 2014) based on machine learning to propose an extractive Arabic summary. In this article, we suggest a new approach for Arabic (multilingual), multi-document text summarization. Based on thematic annotation (segmentation and topic identification) using the combination of Latent Semantic Analysis approach (LSA) and latent dirichlet allocation (LDA), then the presentation of segments of each topic in graphs to eliminate the inter-redundancy problem also to avoid the loss of phrases (sentences) belong to a specific topic. The modified pagerank algorithm is applied to reduce the graphs built and keep only the most important sentences in the final summary. The paper organization is described as section 2 identifies text summarization related works based on topic iden-

Merniz, A., Chaibi, A. and Ben Ghézala, H.
Multi-document Arabic Text Summarization based on Thematic Annotation.
DOI: 10.5220/0010557906390644
In *Proceedings of the 16th International Conference on Software Technologies (ICSOFT 2021)*, pages 639-644
ISBN: 978-989-758-523-4

tification as well as based graph approaches, section 3 presents research goal, section 4 describes the proposed approach for multi-document Arabic summarization, section 5 is dedicated to evaluating proposed approach also to discussing results, the conclusion is presented in section 6.

2 RELATED WORK

There are various automatic text summarization techniques, statistical-based approaches, machine learning-based approaches, linear programming-based approaches, graph-based approaches, and others that combine different techniques to summarize. This section focuses on graph-based approaches as well as topic identification based methods.

Topic identification based approach: (Hennig, 2009) considered a new method for multi-documents summarization to present sentences and queries like probability distributions over latent topics. This approach merges query focused and thematic features computed in the latent topic space to estimate the summary relevance of sentences.

(Hammo et al., 2011) present a hybrid approach for automatic Arabic summarization based on identifying the thematic structure of the input text using a classifier and conceptual thesaurus to select proper sentences gathered from the statistical analysis process.

(Harabagiu and Lacatusu, 2005) proposed a new model for multi-documents summarization based on topic themes using semantic information supplied by semantic parsers. The themes are represented by determining both coherence and cohesion relations that improve the general summary quality.

Graph-based approach:

Extractive Arabic Text summarization of (Elbarougy et al., 2020) bases on graph representation using a pagerank algorithm with a few modifications. Such as the weight of edges between nodes is computed by cosine similarity; also, each node's rank is fixed to the noun number in the sentence; contrary to pagerank, all nodes' rank is equal and fixed to 1/N.

(Mallick et al., 2019) present a summary using modified TexRank algorithm by applying inverse sentence frequency modified cosine similarity To cover the importance of words in sentences also sentences in the document.

(Khan et al., 2018) built a semantic graph for abstractive multi-document summarization such as the nodes represent the predicate-argument structures whereas the edges denote similarity weight. An improved ranking algorithm derives from PageRank is applied to keep only the top-ranked sentences.

(Uçkan and Karcı, 2020) introduce a new method for multi-document summarization based on KUSH algorithm. That prepares data to be presented in graph such as nodes represent sentences and edges are common words. The approach focuses on the concept of removing maximum independent sentences in the graph to preserve only the main nodes in the summary.

3 RESEARCH GOAL

Text summarization is a vast area in Natural Language Processing (NLP). It isn't straightforward to produce a summary system that simultaneously combines several features, especially multi-document, multilingual abstractive text summarization. Besides, when languages are of variable complexity. This research proposes multi-document Arabic text summarization, using a graph to make it flexible in other languages. The graph is independent of the language because graph-based approaches demonstrate an excellent efficiency in automatic text summarization.

4 PROPOSED APPROACH

This paper proposes a hybrid approach for multilingual Arabic multi-document text summarization. Through three phases thematic annotation of documents, graphs representation, graphs reduction and summary construction as shown in algorithm.

Step 1. (Thematic Annotation of Documents): to exploit the thematic aspect of the text, thus facilitating the task of the next phase (graphs representation). In this step, the segmented documents of(Naili et al., 2017) are used who proposed a multilingual analyzer for segmentation and topic identification using semantic knowledge-based on Latent Semantic Analysis(LSA) also (LDA) Latent Dirichlet Allocation. Documents are segmented such as each segment is labeled by its minor and significant topics, in this article, we are interested in essential topics. A document may include several topics, as shown in the figure 1. That represents two segments of text such as the first one identifies topic 4 (Science) as shown in major topic field, whereas the second segement belongs to topic 1 (Health).

Step 2. (Graphs Representation): this phase aims to avoid the inter-redundancy problem known in multi-document text summarization also to cover the segments of all topics identified in the input docu-

Figure 1: Example of thematic annotation of document in arabic.

ments.Segments of each topic are represented by a graph as shown in figure 2 where the vertices define the sentences in Arabic, whereas the edges represent the adjacency relations between sentences.

Step 3. (Graphs Reduction): at the end, a modified pagerank algorithm is applied to the graphs built to reduce them and keep only the most important sentences included in the final summary. PageRank (Page et al., 1999) The pagerank algorithm implemented at stanford university by (Page et al., 1999) and utilized by the search engine of google to determine the significance of webpage using the graph representation. It has different application fields mainly application in search, browsing, and traffic estimation. The concept of pagerank is that a webpage has a high rank if the sum of the ranks of its back-links is high. The formal definition of this algorithm is shown in equation 1. With d is damping factor fixed at 0.85, damping factor, A: the webpage, C (A): is the number of outgoing links of the page A;

$$PR(A) = (1-d) + dx\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

In our case, the graphs are oriented as specified in figure 2, such as the vertices represent the sentences in arabic. In contrast, the arcs represent the adjacency relations between these sentences. After the algorithm's execution with number of iterations (N=100), a score will be carried out to each node that describes the power of this node in the graphs, only nodes related to a high score will be included in the summary.

Modified PageRank Algorithm: The pagerank algorithm is modified, such as the pages are replaced with document sentences. Also, weight is added between edges. It indicates the cosine similarity among sentences. Cosine similarity metric (Han et al., 2012) measures similarity between vectors that represent term or (phrase) frequency in the document. Let x and y, be two vectors for comparison, cosine similarity formula of x, y is:

$$\cos_{sim} = \frac{x.y}{\|x\|\,\|y\|}, \quad (2)$$

Where $\|x\|$, and $\|y\|$:
are euclidean norm of vectors x, y respectively defined as $\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$.

The application of mentioned modifications gives the following PageRank formula:

$$PR_M = (1-d) + d \times \sum_{j \in In(V_i)} \frac{PR_M(V_j) \times \cos_{sim}(V_i, V_j)}{\left|Out(V_j)\right|} \quad (3)$$

Algorithm 1: Proposed approach algorithm.

Input: multi-document multi-topic;
Output: Summary for each topic;
thematic annotation of documents;
ForEach Topic (Major-topic) from D **do**

Create a Graph -Topic (V, E);

ForEach Graph Topic **do**

Calculate Weight W, using Cosine Similarity between sentences (Sentence-A, Sentence-B);
Update Graph-Topic by adding weight between edges such as G' (V, E, W);
Apply Modified PageRank (MPR) on G' with number of iterations (N) ;
Extract Summary for each topic;

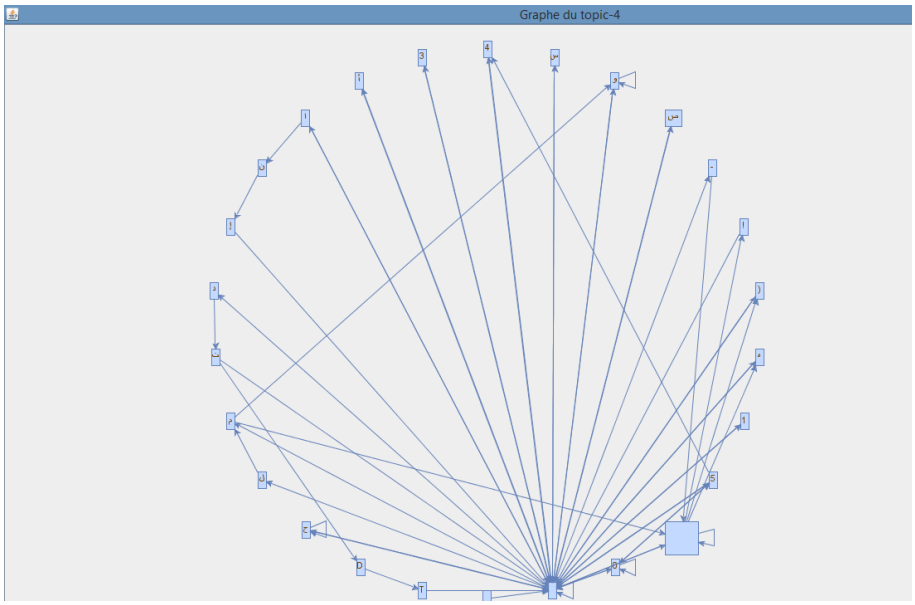

Figure 2: Representation of topic4 (sciences) segments in a graph.

5 EXPERIMENTAL RESULTS AND EVALUATION

5.1 Evaluation Metrics

Summary evaluation is an essential task in measuring and enhancing the results of the summary generated. For this purpose, researchers have developed a variant of metrics that differ according to their mode of application: automatic, semi-automatic, manual also according to their mode of evaluation intrinsic or extrinsic. The extrinsic evaluation concerns the impact assessment of the summary quality on the other tasks such as text classification, information retrieval. Simultaneously, the intrinsic evaluation consists of determining the quality and the informativeness of summary based on reference summaries using automatic metrics or semi-automatic methods. Automatic metrics (Rouge (Lin, 2004) , QARLA (Amigó et al., 2005) , AutoSummENG (Giannakopoulos et al., 2008)) do not involve human annotations, while semi-automatic methods such as relative utility (Radev and Tam, 2003) , factoid score (Teufel and Halteren, 2004), Pyramid (Nenkova and Passonneau, 2004) method involve some of the human annotations.

ROUGE: Recall Oriented Understudy for Gisting Evaluation proposed by (Lin, 2004) to automatically measure the summary quality. It computes the number of units (n-grams, word sequences, and word pairs) common between the summary produced by a machine and a list of reference summaries. ROUGE defines four measures, namely: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S.

Rouge-N: determines the common n-grams between the automatic summary and a list of reference summaries. Rouge-1 for 1-gram, Rouge-2 for bi-grams.

$$Rouge_N = \frac{\sum_{S\in\{RefSum\}}\sum_{gram_n\in\{S\}} Count_{match}(gram_n)}{\sum_{S\in\{RefSum\}}\sum_{gram_n\in\{S\}} Count_{match}(gram_n)} \quad (4)$$

Rouge-L: calculates the maximum length of common sequences (LCS) namely X and Y. such as X represents a collection of sequences from reference summaries whereas Y is a set of sentences of the system summary. (Lin and F.J. Och, 2004) proposes using LCS-based F-measure to estimate the similarity between X and Y of length m and n respectively, where $\beta = \frac{Rouge_L(P)}{Rouge_L(R)}$ according to Eqs :

$$Rouge_L(R) = \frac{LCS(X,Y)}{m} \quad (5)$$

$$Rouge_L(P) = \frac{LCS(X,Y)}{n} \quad (6)$$

$$Rouge_L(F) = \frac{(1+\beta^2)Rouge_L(R)Rouge_L(P)}{Rouge_L(R)+\beta^2 Rouge_L(P)} \quad (7)$$

Rouge-SU4: ROUGE-S: (Skip-Bi-gram Co-Occurrence Statistics) computes the overlap of skip

Table 1: rouge-L, rouge-1 evaluation results of the proposed algorithm.

	Rouge-L			Rouge-1		
	R	P	F	R	P	F
Topic 1 (economics)	0,21126	0,15175	0,17663	0,57128	0,37169	0,45036
Topic 2 (health)	0,14743	0,11429	0,12876	**0,59024**	0,37109	0,45568
Topic 3 (politics)	0,13760	0,28634	0,18587	0,26943	0,52174	0,35535

Table 2: rouge-2, rouge-SU4 evaluation results of the proposed algorithm.

	Rouge-2			Rouge-SU4		
	R	P	F	R	P	F
Topic 1 (economics)	0,35310	0,21996	0,27106	0,39250	0,23860	0,29678
Topic 2 (health)	0,36019	0,21690	0,27076	0,42045	0,24400	0,30879
Topic 3 (politics)	0,12578	0,23393	0,16360	0,15952	0,28675	0,20500

bi-grams between a given summary and its set of references. Skip Bi-gram of a sentence is defined as the couple of words following their same order in the sentence with random gaps. This version has been extended by adding 1-gram as a counting unit to enhance the primary method.

5.2 Corpus

The proposed approach is tested on al sulaiti corpus proposed by (Al-Sulaiti and Atwell, 2006) that contains documents in Arabic divided into several categories(Autobiography, Short Stories, Children's Stories, Economics, Education, Health and Medicine, Interviews, Politics, Recipes, Religion, Sociology, Science, Sports, Tourist, Travel). This study interests in the following subjects: health, politics, economic, and sciences. At the end, each topic obtains its summary system.

5.3 Results Discussion

Tables (1,2) show the summary results of rouge metrics: rouge-L, rouge-1, rouge-2, and rouge-SU4, for each topic. We notice that the results of topics 1 (economics) and 2 (health) are almost similar in rouge-1, rouge-2, and rouge-SU4, with values approximately close to 0.58, 0.36, and 0.41, respectively. However, values of topic 3 (politics) vary between 0.12 in rouge-2 and 0.27 in rouge-1. The proposed approach aims to test the impact of segmented documents combing with the pagerank algorithm on multi-document summarization. It differs from Elbarougy's system and existing studies in the following aspects: first (Elbarougy et al., 2020) suggests arabic summary for single document whereas our approach presents arabic summarization for multi-document. Second the proposed system represents each topic in documents by a graph in order to eliminate the interreduncy problem also it combines a both thematic annotation and pagerank algorithm to produce the summary. (Elbarougy et al., 2020) does not apply annotated documents. Several works highlight graph-based methods, including pagerank (Page et al., 1999) , textrank (Mihalcea and Tarau, 2004) , or lexrank (Erkan and Radev, 2004) algorithms, which have shown excellent efficiency, but those who combine either segmented and Modified pagerank algorithms are limited. Also, we don't use the same corpus. Therefore, we cannot compare to them in this article.

6 CONCLUSION AND FUTURE WORK

Document text summarization is a varied field, rich in characteristics. This paper has introduced the basics definitions of an automatic summary and different related works to text summarization. We proposed Arabic multi-document text summarization appraoch based on segmented multi-topic documents. Separation of each topic in the graph to minimize redundancy also applies to the modified PageRank by adding a cosine similarity measure to the initial PageRank formula. Results of the proposed approach are considerable, with 0.59 (rouge-1) as a high value. For future work, we modify input documents using non-segmented documents to extract the segmentation and topic identification aspect's contribution. We enrich the graph semantically by applying other similarity measures to enhance the quality of the summary system. We make the proposed algorithm multilingual by testing it on different languages such as English, French.

REFERENCES

Al-Sulaiti, L. and Atwell, E. (2006). The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

Amigó, E., Gonzalo, J., Penas, A., and Verdejo, F. (2005). Qarla: a framework for the evaluation of text summarization systems. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 280–289.

Azmi, A. and Al-Thanyyan, S. (2012). A text summarizer for arabic. *Computer Speech & Language*, 26(4):260–273.

Conroy, J., Davis, S., Kubina, J., Liu, Y., O'leary, D., and Schlesinger, J. (2013). Multilingual summarization: Dimensionality reduction and a step towards optimal term coverage. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 55–63.

El-Haj, M. and Rayson, P. (2013). Using a keyness metric for single and multi document summarisation. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 64–71.

Elbarougy, R., Behery, G., and Khatib, A. E. (2020). Extractive arabic text summarization using modified pagerank algorithm. *Egyptian Informatics Journal*, 21(2):73–81.

Erkan, G. and Radev, D. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Fejer, H. and N.Omar (2014). Automatic arabic text summarization using clustering and keyphrase extraction. In *Proceedings of the 6th International Conference on Information Technology and Multimedi*, pages 293–298. IEEE.

Giannakopoulos, G., Karkaletsis, V., Vouros, G., and Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):1–39.

Hammo, B., Bassam, H., h. Abu-Salem, and Evens, M. (2011). A hybrid arabic text summarization technique based on text structure and topic identification. *International Journal of Computer Processing of Languages*, 23(01):39–65.

Han, J., Kamber, M., and Pei, J. (2012). 13-data mining trends and research frontiers. *Data Mining (Third Edition), ed Boston: Morgan Kaufmann*, pages 585–631.

Harabagiu, S. and Lacatusu, F. (2005). Topic themes for multi-document summarization. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 202–209.

Hennig, L. (2009). Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of the International Conference RANLP-2009*, pages 144–149.

Khan, A., Salim, N., Farman, H., Khan, M., Jan, B., Ahmad, A., Ahmed, I., and Paul, A. (2018). Abstractive text summarization based on improved semantic graph approach. *International Journal of Parallel Programming*, 46(5):992–1016.

Lin, C. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, C. and F.J. Och, F. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Mallick, C., Das, A., Dutta, M., Das, A., and Sarkar, A. (2019). Graph-based text summarization using modified textrank. In *Soft computing in data analytic*, pages 137–146. Springer.

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

Naili, M., Chaibi, A., and Ghézala, H. (2017). Arabic topic identification based on empirical studies of topic models. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées*, 27.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Radev, D. and Tam, D. (2003). Summarization evaluation using relative utility. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 508–511.

Teufel, S. and Halteren, H. (2004). Evaluating information content by factoid analysis: human annotation and stability.

Uçkan, T. and Karcı, A. (2020). Extractive multi-document text summarization based on graph independent sets. *Egyptian Informatics Journal*, 21(3):145–157.