Optimization of Multi-Document Summarization through Multi Agent Systems

Riva Malik
FAST-NUCES
Islamabad, Pakistan
i201208@nu.edu.pk

ABSTRACT

Automatic text summarization aims to achieve non-redundancy, document coverage and relevancy. Considering these features the process of optimal sentence selection for summarization becomes a NP-Hard problem. In order to solve this problem we apply meta-heuristic optimization algorithms to obtain optimal summaries. Existing approaches utilize these algorithms for single document summarization. In our proposed notion, we extend this optimization problem to multi-document summarization by applying the concepts of multi agent systems. Each document is passed to an agent where the summary optimization takes place locally for single document sentence selection. Furthermore these each agent cooperates to produce multi-document summary using the sentences generated by each agent. Those sentences are considered for summary that remain after the optimization process is completed.

KEYWORDS

Text Summarization, Multi Agent Systems, Optimization

ACM Reference Format:
Riva Malik. 2023. Optimization of Multi-Document Summarization through Multi Agent Systems. In *ACM Conference, Washington, DC, USA, July 2017*, IFAAMAS, 3 pages.

1 INTRODUCTION

There has been an exponential increase in sources of online information like news, articles and blogs. This has made it difficult for user to obtain correct information in shorter time. The user requires the information to be presented in concise yet comprehensive form. For this purpose automatic text summarization which utilizes computing algorithms to obtain summary, has gained popularity.

The ultimate goal of automatic text summarization is to achieve a summary the fulfills the following criteria [9]:

- Non- redundancy: The summary should not contain repeated information.
- Coverage: All the salient information present in document should become part of the summary and no important information should be omitted.
- Relevancy: The information relevant to the user should be the part of summary

A summary is considered optimum if it has the above mentioned qualities.

Automatic text summarization is divided into different categories: based on the number of input document it is identified as single document and multi-document summary. Considering the methodology it is categorized as extractive and abstractive summarization. Extractive summarization approaches extract the salient textual units like terms, sentences or paragraphs and utilize them without any modification in the resultant summary.

Single document extractive summarization utilizes various approaches namely unsupervised learning, supervised learning and meta heuristic algorithms. Due to the features of an ideal summary and the variability in comparative parameters of information the task of summarization becomes an NP-Hard problem. Apart from that large volume of information and search space of textual units makes the sentence selection difficult. Hence meta heuristic can be applied to achieve optimal solution to summarization problem [8]. There are several optimization algorithms like Genetic Algorithm(GA), Particle Swarm Algorithm(PSO) and Firefly algorithm [1] used to optimize single document summarization problem.

With the recent advancement in distributed artificial intelligence the optimization algorithms are used along with the concepts of multi agents systems to get efficient solutions. These multi agents possess the following properties:

- Residing in an environment called world view
- Ability to make intelligent decisions
- Having reactive behaviour based on the other agents in the environment

In this paper we propose a notion to solve the multi-document summarization problem using meta heuristic algorithm based in a multi agent system.

2 CONCEPTS AND DEFINITIONS

This section presents the concepts along with the relevant literature of the domain.

2.1 Text Summarization

Text summarization is the process of extracting important information from large collection of text to present it in a concise form. The objective and approach of summarization is used as a criteria to divide it into various categories. Depending on the number of documents to be summarized, it is classified into single document and multi document [7]. Single document summarization considers

ACM Conference, , July 2017, Washington, DC, USA. © 2023 Association for Computing Machinery. . . . $ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
. . . $15.00

one document for extraction of salient sentences to form summary whereas in multi document summarization various related documents are used to form a summary that captures the important information present in all of the given documents. The summarization approach is mainly divided into two categories namely extractive summarization and abstractive summarization [5]. Extractive summarization selects the relevant sentences present in source documents without any modification and combines them to form the summary. Abstractive summarization captures the important information from the source documents and reconstructs the sentences using deep learning techniques to form the summary. These summarizers have better ability to capture the semantic structure source documents [10].

2.2 Extractive Text Summarization

Extractive text summarization is achieved using unsupervised learning,supervised learning and meta heuristic techniques. Unsupervised learning based methods generally generate summary on the fly as they do not require any training on the source documents. These methods utilize statistical features of the documents [2]. Documents are organized in such a way that they cover information on various topics in sequential manner, using this property the sentences of the documents can be clustered into various topics and the sentences that are to be included in the summary are chosen from these clusters [6]. This technique provides better document coverage as compared to statistical methods.

2.3 Meta Heuristic Techniques

Meta heuristic techniques provide a set of optimization algorithms used for solving the extractive summarization problem. In this category the summarization task is considered as NP-Hard problem due to variability in parameters that are used to achieve non-redundancy, coverage and relevancy. [8] in their approach present a combination of biogeography-based optimization (BBO) algorithm and multi-agent systems concepts to generate an optimum summary. [3] uses GA whereas [4] employ particle swarm optimization for selection of sentences to be included in summary.

2.4 Firefly Algorithm

Firefly algorithm is a meta heuristic, population based algorithm. It takes inspiration from the flashing nature of fireflies. This algorithm works on the principle that fireflies are attracted to each other, where the attraction is propotional to brightness. This means that less bright fireflies are attracted to brighter fireflies. As the distance between the fireflies increases the attraction decreases. If the brightness has the same values then the fireflies can move randomly. The algorithm is defined in the following steps:

(1) The population is randomly initialized
(2) Calculate brightness and movement of each firefly
(3) Update the brightness and position of each firefly
(4) Check for stopping criteria, if it is met then the fireflies are ranked to return the optimal solution

3 PROPOSED NOTION

The proposed solution consists of the following steps:

(1) Document pre-processing
(2) Score calculation
(3) Firefly algorithm application in multi agent environment

3.1 Text Pre-processing

In order for any algorithm to work we have to convert the text into the form understandable by the computer that is why preprocessing is a vital step in any Natural Language Processing(NLP) task. The preprocessing tasks involve word tokenziation, sentence segmentation, stop word removal, special character removal, stemming and lemmatization to name a few. In our approach we apply the below mentioned pre-processing techniques:

- Sentence Segmentation: The document is divided into sentences on the basis of fullstop.
- Stop word removal: Stop words do not play any vital role in in candidate sentence selection because they are repeated in every sentence. In fact they can further be the cause of redundancy in summary, hence they are removed from the sentences
- : Lemmatization: Lemmatization converts the word into its base form. This step is important because many words that we speak have the same base form hence they should be treated at same level especially when we have to apply statistical methods on the terms and sentences.
- Special Character Removal: Special characters hinders the process of summarization hence they are removed.

3.2 Score Calculation

Structural and statistical features of the documents are considered to to extract the relationship between sentences. The relationship will help us in generating candidate sentences for the summary. We calculate the following values as scores.

3.2.1 Sentence Length(SL). The length of sentence(S) is calculated by the formula:

$$SL = \frac{No\ of\ words\ in\ S}{No\ of\ word\ in\ longest\ S} \quad (1)$$

3.2.2 Sentence Position(SP). The sentence position(SP) is defined as the location of sentence occurrence within the a single document. It is calculated by the formula:

$$SP = 1 - \frac{SP\ of\ S}{Total\ Sentences} \quad (2)$$

3.2.3 TF-IDF. TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a statistical method used in NLP to find out the relevance of a term to a document. It is also used to find out important terms in a document.

We calculate the TF-IDF of term t in sentence S where TF is the number of times the term occurs in single sentence. IDF is calculated using the formula:

$$IDF(t) = \frac{log(N)}{log(N_t)} \quad (3)$$

where N is the total number of sentences and N_t is the number of sentences in which t occurs.

Now we combine all the above calculated scores to form a collective statistical score using formula:

$$TS = SL + SP + TF - IDF \quad (4)$$

3.2.4 Cosine Similarity. Apart from statistical scores we also calculate the semantic similarity between the sentences using cosine similarity.

3.2.5 Document Similarity Score. Apart from calculating the similarity between the sentences within the document. We also compute the inter document similarity. Lets consider there are two documents D_1 and D_2. In the first step we sum up the *TS* of each sentence in each document. Then the difference between the sum of *TS* is calculated and scaled between 0 and 1. We assume that document similarity increases when the score is greater than 0.5 and decreases vice versa.

3.3 Firefly Algorithm Application in Multi Agent Environment

In this section we apply the Firefly algorithm for multi document summarization using the concepts of multi agent systems. We consider the number of agents equal to the number of documents for summarization, in other words we assign each document to each agent. The agents work locally and cooperate globally to optimize the summarization process. Each agent applies the Firefly algorithm internally to generate sentences for candidate summary from that particular document. Now each agent outputs a set of candidate sentences which we consider as documents. In global level we compute the similarity between the set of candidate sentences generated by each agent and use this similarity as feedback. If the document similarity is greater we provide the feedback to those particular agents to remove those sentences from the candidate sentences in the next round of Firefly iteration. In this way we aim to reduce the redundancy in the summary meanwhile ensuring better coverage. This process is continued until the stopping condition of Firefly algorithm is met and there remains similarity less than 0.5 between all the sentences generated by the agents. These sentences are categorized as final summary.

4 CONCLUSION

Automatic text summarization employs the aid of computing algorithms to obtain summaries. The task of sentence selection in the process of summarization is a difficult task making it an NP-Hard problem. This problem can be solved using optimization algorithms which results in optimal selection of sentences in the summary. Existing approaches apply these meta heuristic algorithms to obtain summary from a single document. We propose an approach that performs multi document summarization by using Firefly meta heuristic algorithm. Apart from that we employ the algorithm in multi agent environment where each agent computes candidate sentences and cooperate with one another to generate optimal summary at multi document level. In future we can implement this work to obtain concrete results from our proposed notion.

REFERENCES

[1] Raed Z Al-Abdallah and Ahmad T Al-Taani. 2019. Arabic text summarization using firefly algorithm. In *2019 amity international conference on artificial intelligence (AICAI)*. IEEE, 61–65.

[2] Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. 2016. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications* 7, 4 (2016), 285–294.

[3] Mohamed Abdel Fattah and Fuji Ren. 2009. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech & Language* 23, 1 (2009), 126–144.

[4] Oi-Mean Foong and Alan Oxley. 2011. A hybrid PSO model in extractive text summarizer. In *2011 IEEE Symposium on Computers & Informatics*. IEEE, 130–134.

[5] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[6] Kalliath Abdul Rasheed Issam, Shivam Patel, et al. 2021. Topic modeling based extractive text summarization. *arXiv preprint arXiv:2106.15313* (2021).

[7] Rada Mihalcea and Paul Tarau. 2005. A language independent algorithm for single and multiple document summarization. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

[8] Seyed Hossein Mirshojaee, Behrooz Masoumi, and Esmaeil Zeinali. 2020. Mamhoa: a multi-agent meta-heuristic optimization algorithm with an approach for document summarization issues. *Journal of Ambient Intelligence and Humanized Computing* 11, 11 (2020), 4967–4982.

[9] Houda Oufaida, Omar Nouali, and Philippe Blache. 2014. Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. *Journal of King Saud University-Computer and Information Sciences* 26, 4 (2014), 450–461.

[10] Tham Vo. 2021. SE4ExSum: An Integrated Semantic-aware Neural Approach with Graph Convolutional Network for Extractive Text Summarization. *Transactions on Asian and Low-Resource Language Information Processing* 20, 6 (2021), 1–22.