# Exploring Multi Feature Optimization for Summarizing Clinical Trial Descriptions

Saichethan Miriyala Reddy
*Indian Institute of Information Technology, Bhagalpur*
*Email: miriyala.cse.1725@iiitbh.ac.in*

Saisree Miriyala
*Adobe Inc.*
*Email: mirilaya@adobe.com*

***Abstract*—Documenting Clinical Trial Descriptions of patients can help doctors with diagnostics and treatment plans and can be used for future reference. However, with the rapid growth of population, manually checking all previous files of a patient is not feasible. We address this challenge by providing summaries of clinical trial descriptions. We present a framework for automatically summarizing Clinical Trial Descriptions, which takes advantage of different features in semantic and syntactic space. We propose a multi objective optimization technique which uses position and similarity of the sentences. The similarity is established based on TF-IDF and WMD. We evaluate the proposed method on clinical trial dataset, and compare the results against human gold standard summaries using ROUGE metrics. Our approach is unsupervised in nature which has an advantage over supervised models in the advent of new diseases where a large quantity of quality data is not available. We provide a detailed ablation study to show the contribution of each feature in our approach and release our code on GitHub.**[1]

***Keywords*-summarization ; multi objective; clinical trials;**

## I. Introduction

The recent outbreak of COVID-19 has catalyzed the development of novel coronavirus vaccines. However, the creation of any vaccine or drug takes much time because of the multiple steps involved, and one of the significant steps after the development of the vaccine is clinical trials. Clinical trials are a type of research that studies new tests and treatments and evaluates their effects on human health. Clinical Trial summarization is a task of creating salient, non-redundant and condensed variants of trial descriptions. Extractive summarization methods using sentence features are popular. The informational sentences are selected from long descriptions using sentence scoring features. Some of the sentence scoring features are sentence length, position, similarity and dissimilarity with title and Maximum Marginal Relevance(MMR)[1]. It has been shown in the literature survey that the chances of a sentence being in the generated summary is higher if the sentence is longer, placed closer to the title, and with less word mover distance (WMD) from the title than the other sentences.

Feature-based summarisation has shown consistently reliable performance at summarising documents. In this paper, a novel summarization framework is developed employing the concept of multi-feature optimization. Firstly, clinical trial descriptions are split into a list of sentences. The sentences are ranked based on different features like position and similarity with title in semantic and syntactic spaces. In this current work, we describe a multi-feature based clinical trial description summarization framework that generates brief, non-redundant synopses from long-form trial descriptions. To evaluate the quality of our generated summaries, ROUGE scores between human summary and system summary are used. Note that features like length of sentences and MMR are not considered in this approach because most sentences in trial descriptions have similar length and the number of sentences is typically too low for effective results.

[1]https://github.com/Saichethan/MFOS

## II. Contributions

The major contributions of the current paper are enumerated below:

1) A Multi-Feature Optimization framework is proposed for summarizing clinical trial descriptions by considering different semantic features.
2) Ablation study is performed to analyze the most contributing feature for summary generation.

## III. Background

### A. Related work

In this section we will briefly review existing unsupervised algorithms in the context of extractive text summarization tasks.

**LexRank**[2] and **TextRank**[3] are stochastic graph-based approaches which compute the importance of sentences based on centrality of eigenvector in a graph representation of sentences in a document or group of documents. it measures the agreement of each sentence in a given cluster and extracts the most crucial sentence to include in the generated summary. The key difference between the two approaches is TextRank assumes all weights to be unit weights, whereas LexRank uses a weighted function for assigning weights.

**LSA**[4] uses the relevance score between the weighted term frequency vector of each sentence and the weighted term frequency of the document. It assumes that sentences

978-1-7281-9325-0/20/$31.00 ©2020 IEEE
DOI 10.1109/BigMM50055.2020.00059

with a higher relevance score have a higher probability for inclusion in summary.

**Luhn**[5] ranks the sentences based on keyword frequency and proximity within a sentence. Moreover, sentence weights are determined by looking for significant words in a sentence.

**SumBasic** [6] produces general multi document summaries; it is based on the intuition that words occurring more frequently in the document or sentence cluster occur with the higher possibility in the human gold summaries than words occurring less frequently.

**KLSum**[7] is a greedy method, in which the goal is to find a set of summary sentences which closely match the document set unigram distribution.

### B. Evaluation Metrics

For evaluating our results, i.e. quality of generated summaries, we use ROUGE[8] metrics. There are different variants (ROUGE-N, ROUGE-L, ROUGE-S) [8][9] which will be discussed briefly here.

**ROUGE-N** is recall and precision oriented metric which considers precision or recall of N-grams in the summary generated to gold standard summary. In N-gram metric N usually ranges from 1 to 4.

**ROUGE-L** compares system generated summary with respect to gold standard based on the Longest Common Subsequences (LCS). The longer the common sequences between the system and human summaries the more will be the similarity and the importance of the sequences in the generated summary.

**ROUGE-S** calculates the skip-bigram co-occurrence statistics between the generated and gold summaries. ROUGE-2 is comparable to ROUGE-S except that ROUGE-2 does not support gaps between the bigrams.

For our experiment we used ROUGE-1, ROUGE-2 and ROUGE-L. They are the most commonly used variants for evaluating summarization tasks since they show high correlations with human judgment of the quality of summaries. Among these three, ROUGE-L measures the coherence of sentences better than others.

## IV. Methodology

### A. Problem definition

Consider a clinical trial description $TD$ which consists of $N$ sentences, $s_1$, $s_2$, ..., $s_N$. Our goal is to find the subset of the sentences, $S \epsilon TD$, which encapsulates the essence of the original trial description and satisfies length constraints.

$$\sum_{s_i \epsilon S} l_i \leq L_{limit} \tag{1}$$

where $S$ represents the gist of the clinical trial description , $s_i$ is a sentence belonging to S, and $l_i$ measures the length of $i^{th}$ sentence in terms of number of words, $L_{limit}$ is the word limit of summary. In our experiments we set $L_{limit}$ as $1/3$ rd of original description length.

### B. Proposed Approach

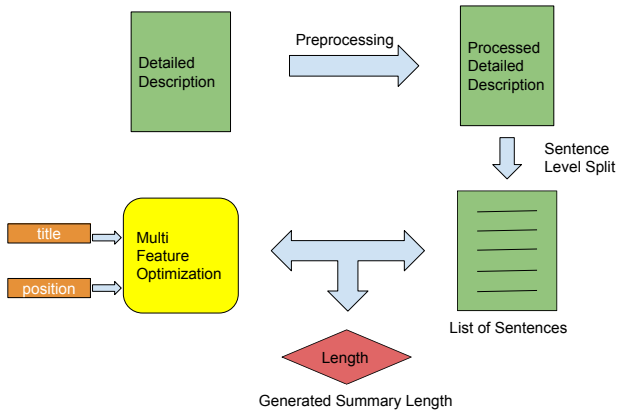


Figure 1. Flow Chart of the proposed architecture

For summary generation, the following steps are considered:

- Firstly break down the detailed description into list of sentences[2]. Let $N$ be the number of sentences in a clinical trial description.
- Calculate the sentence scores of all the sentences using the following features: sentence position[10], similarity with title[11], word mover distance with title. We discuss these features in more detail here:
  1) Sentence position: The starting sentences or paragraphs of the document are most important to us as they tend to convey relevant information [10] [12]. So, the lower is the index value of the sentence in the document, the more important it will be.

$$F1 = \sqrt{\frac{1}{p+1}} \tag{2}$$

  Where $p$ is the position of the sentence in the detailed description
  2) Similarity of sentences with the title of the clinical trial using cosine similarity of their TFIDF vector.

[2]Preprocessing step include removal of unwanted characters and symbols

The sentences which are more similar to the title in semantic space are more probable to appear in the final summary [13]

$$F2 = \frac{c_i}{max(c_j)} \tag{3}$$

$$0 \leq j \leq N \tag{4}$$

Where $c_i$ is the cosine similarity of ith sentence with detailed description title. In general for two vectors A and B cosine similarity is defined as

$$CosineSimilarity(A, B) = \frac{A.B}{|A| \times |B|} \tag{5}$$

3) Similarity of sentences with the title of the clinical trial descriptions using word mover distance. It has been shown that the smaller the word mover distance between a sentence and the title the more likely the sentence is to appear in the final summary [10]

$$F3 = \sqrt{\frac{1}{d_i + 1}} \tag{6}$$

Where $d_i$ is the word mover distance between the ith sentence in detailed description with the title.

Using these three sentence scoring functions, the cumulative score of the ith sentence in the detailed description can be calculated as:

$$MFO_i = \alpha \times F1 + \beta \times F2 + \gamma \times F3 \tag{7}$$

Where,

$$\alpha + \beta + \gamma = 1 \tag{8}$$

$$0 \leq F1, F2, F3 \leq 1 \tag{9}$$

We have experimented with three variants of Multi Feature Optimization (MFO): MFO-P, MFO-T, MFO-W (defined in Table II). The final objective function is chosen as the convex optimization of these variants.

$$MFO{-}PTW = Max_{ROUGE-L}(MFO{-}P, MFO{-}T, MFO{-}W)$$

This score is used as the objective function to extract the most informative sentences until the desired summary length is reached. The flow chart of the proposed architecture is as shown in figure1.

**Algorithm 1:** Multi-Feature Optimization (**MFO**)

```
Split the trial description into individual sentences;
initialize hyperparameters based on Table II;
initialize summary = [ ];
sort sentences in trial description, based on scores of
  objective function;
while len_summary < L_limit do
    if two or more sentences have same score then
        if if this score is zero then
            boost the score of sentence closest to the
              title
        end
        else
            boost the score of sentence with most
              word overlap;
        end
    end
    add top sentence to summary;
    remove top sentence from candidate set;
    if candidate set is empty then
        break;
    end
end
Result: most relevent sentences in the given
        candidate set
```

## V. Experiments

In this section, we first present and describe the dataset used in our experimental setup. We then evaluate our approach comparing with existing methods.

### A. Dataset

Data for extractive summarization of clinical trial descriptions[14] is publicly available in Mendeley datasets[3]. It consists of 101016 clinical trials. Each clinical trial instance has ten attributes out of which we have considered four for our experiment. They are NCT ID which is a unique ID of trial, title, detailed description, and brief summary which is considered as gold summary. Brief statistics of the used dataset are given in Table I.

Table I
Brief description of dataset used

| | Min | Mean | Max | Std Dev |
|---|---|---|---|---|
| Desc. len in words | 10 | 334.8 | 5110 | 358.2 |
| Summary len in words | 2 | 86.76 | 870 | 81.17 |
| Desc. len in Sentences | 2 | 14.02 | 302 | 15.64 |
| Summary len in Sentences | 1 | 3.383 | 61 | 3.465 |

Hyperparameters used in different variants of our approach are shown in Table II

### B. Comparative Methods

We compare our results with those of existing methods like LexRank[2], TextRank[3], LSA[4], Luhn[5],

[3]http://dx.doi.org/10.17632/gg58kc7zy7.1file-46bd6ba5-0991-4929-9588-0a40dcad16d3

Table II
HYPERPARAMETERS USED

| Variant | Values |
|---|---|
| **MFO-P** | $\alpha = 1; \beta = 0; \gamma = 0$ |
| **MFO-T** | $\alpha = 0; \beta = 1; \gamma = 0$ |
| **MFO-W** | $\alpha = 0; \beta = 0; \gamma = 1$ |

| | ROUGE 1 | | | ROUGE 2 | | | ROUGE L | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***P*** | ***R*** | ***F*** | **P** | ***R*** | ***F*** | ***P*** | ***R*** | ***F*** |
| Random | 0.31 | 0.30 | 0.29 | 0.13 | 0.12 | 0.12 | 0.27 | 0.26 | 0.2508 |
| LexRank[2] | **0.35** | 0.35 | 0.34 | **0.17** | 0.17 | 0.16 | **0.31** | 0.32 | 0.29 |
| TextRank[3] | 0.34 | 0.38 | 0.35 | 0.16 | 0.18 | 0.17 | 0.30 | 0.33 | 0.30 |
| LSA[4] | 0.33 | 0.36 | 0.34 | 0.16 | 0.17 | 0.16 | 0.29 | 0.32 | 0.29 |
| Luhn[5] | 0.34 | 0.37 | 0.34 | 0.16 | 0.18 | 0.16 | 0.30 | 0.33 | 0.29 |
| SumBasic[6] | 0.33 | 0.29 | 0.30 | 0.13 | 0.12 | 0.12 | 0.29 | 0.26 | 0.25 |
| KLSum[7] | 0.32 | 0.31 | 0.31 | 0.14 | 0.14 | 0.13 | 0.28 | 0.28 | 0.26 |
| **MFO-PTW** | 0.27 | **0.70** | **0.37** | 0.15 | **0.39** | **0.20** | 0.19 | **0.50** | 0.26 |

Table III
COMPARISON BETWEEN PROPOSED METHOD AND EXISTING METHODS

SumBasic[6] and KLSum[7]. More details about these methods are provided in **Related work** section.

## VI. RESULTS AND DISCUSSION

### A. Comparison with existing methods

The comparison among the best ROUGE scores obtained using **MFO** and existing methods is given in Table III. Here

$$Precision(P) = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Recall(R) = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$F - Score(F) = 2 * (\frac{Precision * Recall}{Precision + Recall})$$

From this table, it can be analyzed that our F-scores of ROUGE-1 and ROUGE-2 have an improvement of $5\%$ and $16\%$ respectively. However F-score of ROUGE-L is less by $14\%$. This can be explained by the fact that our approach does not consider the order of sentences. Note that our model has higher Recall values than any other approach, which means our approach has very low false-positive cases. We observed that when length of gold summary is much less than $1/3^{rd}$ of detailed description length often resulted in bad generated summaries. Two of the best system generated summaries are shown in the Table IV.

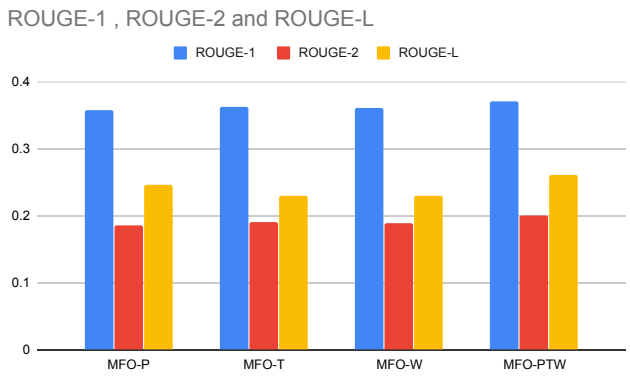


Figure 2. Avg. F Scores of different variants

### B. Comparison among different variants

The results obtained using different variants of the proposed approach are shown in Figure 2. The highest ROUGE-1 and ROUGE-2 scores are 0.36 and 0.19 respectively for the **MFO-T** variant, which considers cosine similarity of sentences with the title of the trial.

## VII. CONCLUSION AND FUTURE SCOPE

We propose a multi-feature based optimization model **MFO** for summarization task, where the different features like position, cosine similarity with title, and word mover distance are used. In addition to these features we can also consider using features like length and maximum marginal relevance [10]. Using a multi-feature clustering approach for summarization on this dataset is yet to be explored. Most of the existing approaches are graph based and unsupervised as the number of medical datasets has rapidly increased in the last decade. The availability of diverse datasets has opened the doors for deep learning based approaches in the future. However one of the major concerns with such approaches is privacy[15], but recent methods like federated learning can be employed in this regard.

## REFERENCES

[1] J. F. Forst, A. Tombros, and T. Roelleke, "Less is more: Maximal marginal relevance as a summarisation feature," in *Conference on the Theory of Information Retrieval*. Springer, 2009, pp. 350–353.

[2] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[3] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

Table IV
THIS TABLE CONTAINS SOME OF THE BEST SUMMARIES GENERATED USING **MFO**

| |
|---|
| **NCT02970812**<br><br>**System Summary:** Background: This randomized, placebo-controlled, double-blind, controlled study was designed to investigate the efficacy of electrical muscle simulation (EMS) for treatment of waist circumference (WC) reduction in abdominal obese adultsMethods: 60 patients with abdominal obese, man with WC ¿ 90 cm and woman with WC ¿ 80 cm, received EMS as experimental group (EG) or transcutaneous electrical nerve stimulation (TENS) as control group (CG) 5 times a week for 12 weeks.<br><br>**Gold Summary:** This randomized, placebo-controlled, double-blind, controlled study was designed to investigate the efficacy of electrical muscle simulation (EMS) for treatment of waist circumference (WC) reduction in abdominal obese adults" "60 patients with abdominal obese, man with WC ¿ 90 cm and woman with WC ¿ 80 cm, received EMS as experimental group (EG) or transcutaneous electrical nerve stimulation (TENS) as control group (CG) 5 times a week for 12 weeks. |
| **NCT02430090**<br><br>**System Summary:** 45 pregnant women undergoing cesarean section were enrolled in the study in november 2006 to march 2007Using CSE technique, 2 ml of 0.5 levobupivacaine was added to 1 ml of saline in group I, 1 ml of 15 mu cg of fentanyl in group II and 1 ml of 1,5 mu cg sufentanil in group III by intratechal administrationHemodynamic parameters, characteristics of sensory and motor blockade, peri-operative and postoperative visual analogue scale (VAS) pain scores, the time to the first analgesic requirement and adverse effects were recorded.<br><br>**Gold Summary:** 45 pregnant women undergoing cesarean section were enrolled in the study in november 2006 to march 2007 "2 ml of 0.5 levobupivacaine was added to 1 ml of saline in group I, 1 ml of 15 mu cg of fentanyl in group II and 1 ml of 1,5 mu cg sufentanil in group III by intratechal administration" "Hemodynamic parameters, characteristics of sensory and motor blockade, peri-operative and postoperative visual analogue scale (VAS) pain scores, the time to the first analgesic requirement and adverse effects were recorded." |


[4] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 19–25.

[5] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[6] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing & Management*, vol. 43, no. 6, pp. 1606–1618, 2007.

[7] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 362–370.

[8] F. Liu and Y. Liu, "Correlation between rouge and human evaluation of extractive meeting summaries," in *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 2008, pp. 201–204.

[9] A. Cohan and N. Goharian, "Revisiting summarization evaluation for scientific articles," *arXiv preprint arXiv:1604.00400*, 2016.

[10] N. Saini, S. Saha, A. Jangra, and P. Bhattacharyya, "Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm," *Knowledge-Based Systems*, vol. 164, pp. 45–67, 2019.

[11] D. Lawrie, W. B. Croft, and A. Rosenberg, "Finding topic words for hierarchical summarization," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 349–357.

[12] N. Saini, S. Saha, and P. Bhattacharyya, "Multiobjective-based approach for microblog summarization," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1219–1231, 2019.

[13] M. Litvak, M. Last, and M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 927–936.

[14] C. Gulden, M. Kirchner, C. Schüttler, M. Hinderer, M. Kampf, H.-U. Prokosch, and D. Toddenroth, "Extractive summarization of clinical trial descriptions," *International journal of medical informatics*, vol. 129, pp. 114–121, 2019.

[15] S. M. Reddy and S. Miriyala, "Security and privacy preserving deep learning," *arXiv preprint arXiv:2006.12698*, 2020.