Multi-document Summarization using Evolutionary Multi-objective Optimization

Chihoon Jung
KSE, KAIST, South Korea
chihoon.jung@kaist.ac.kr

Rituparna Datta
KSE, KAIST, South Korea
rdatta@kaist.ac.kr

Aviv Segev
KSE and CS, KAIST, South Korea
aviv@kaist.edu

ABSTRACT

Text summarization aims to generate condensed summary from a large set of documents on the same topic. We formulate text summarization task as a multi-objective optimization problem by defining information coverage and diversity as two conflicting objective functions. The result solutions represent summaries that ensure the maximum coverage of the original document and the diversity of the sentences in the summary among each other. The initial experiment using DUC2002 multi-document summarization task dataset and ROUGE evaluation metric shows that the proposed method generates high ROUGE score summaries and is comparable to the state-of-the-art summarization methods.

CCS CONCEPTS

•Information systems →Summarization; •Applied computing →*Multi-criterion optimization and decision-making;* **•Theory of computation →**Bio-inspired optimization;

KEYWORDS

Text Summarization, Evolutionary Multi-objective Optimization

1 INTRODUCTION

In the era of information overload, it is helpful to consume summarized information. We propose a new approach for Multi-document Summarization using Evolutionary Multi-objective Optimization (MS-EMO) to generate extractive text summary. The proposed approach tackles the summarization process in a global evaluation perspective by evaluating all the sentences as a whole for the objective function evaluation where as local evaluation approach evaluates one sentence at a time to evaluate its objective function value which may be trapped in the local optima. Another advantage of the proposed approach is that it produces multiple optimal solutions. We can apply data analysis techniques on the results to select the solution closest to the human generated summary among the non-dominating solutions. We use DUC2002 multi-document summarization task dataset for the evaluation of the method. The task is defined as generic extractive multi-document summarization task. For the query-based summarization task, a query is provided to be used as additional information for the summarization, whereas generic summarization task works without any additional information other than the original document set. Extractive summarization task is to select the sentences from the original document set and compose a summary, whereas an abstractive summarization task allows to modify original content or generate the terms.

We propose formulation of the multi-document summarization task by defining the two objective functions coverage and diversity to be maximized and suggest an algorithm to evaluate both objective functions using text mining techniques. We define a constraint for the summary to include maximum of 200 words for the automatic evaluation using ROUGE metric.

2 METHODOLOGY

MS-EMO consists of two modules that communicate among them throughout the optimal summary generation process. Figure 1 explains the working principle of MS-EMO. The text processing module works on data analysis using diverse text mining techniques to calculate the values of the two objective functions. The optimization module is responsible for generating and optimizing candidate summaries represented as chromosomes in evolutionary algorithms using a multi-objective evolutionary algorithm.

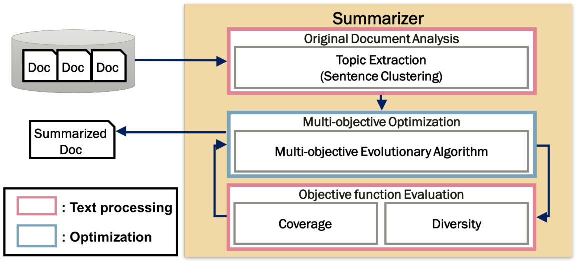
Summarizer
Original Document Analysis
Topic Extraction (Sentence Clustering)
Doc
Doc
Doc
Summarized Doc
Multi-objective Optimization
Multi-objective Evolutionary Algorithm
Objective function Evaluation
Coverage
Diversity
: Text processing
: Optimization

Figure 1: Overview of the MS-EMO

2.1 Text Processing

First, we apply normalization on every term as a preprocessing step. We divide the process into syntactic normalization and semantic normalization. In the syntactic normalization step, we convert the text into lower case letters followed by stemming process. In our work, we use Porter stemming [4] algorithm which is widely used. By going through the stemming, we can successfully remove unnecessary affixes of the terms and match the same terms with diverse derivational affixes. In the semantic normalization step, we add synonyms of the terms in a given sentence. Synonyms are found using WordNet [3] which is a human generated dictionary of word relations for automatic text processing purpose. After applying the preprocessing steps, the text processing module provides two phases. In the first phase, topics are extracted from the set of documents on which automatic summarization is to be performed. To extract the topics, k-means clustering algorithm is applied and centroids of each cluster is used as topics represented as c_i. The

GECCO '17 Companion, Berlin, Germany
© 2017 Copyright held by the owner/author(s). 978-1-4503-4939-0/17/07...$15.00
DOI: http://dx.doi.org/10.1145/3067695.3082040

centroid c_i is calculated as the average of sentence vectors in each cluster C_i. These topics are used in second phase to calculate coverage of candidate summaries returned from optimization module explained in the next subsection. In second phase, binary encoded candidate summaries are returned from the optimization module and passes the calculated value of two objective functions back to the optimization module. This process is iterated until the optimization meets a predefined number of generations as a stop criterion. For the objective function calculation in the second phase, similarities of text pairs (sentences) need to be calculated. This requires the sentences to be represented in a computationally feasible form. We use Vector Space Model (VSM) for the sentence representation. Each term element in a sentence is weighted using logarithm of the term frequency (*tf*) weighting scheme described as w_{ik} in Eq. 1 which denotes kth element in ith sentence in the collection.

$$s_i = \{w_{i1}, w_{i2}, ..., w_{im}\}, \quad w_{ik} = log(tf_{ik}) \tag{1}$$

Once sentences are represented using VSM in m-dimensional space, cosine similarity (Eq. 2) is applied on a pair of sentences represented as s_i and s_j, which denotes ith and jth sentence in the document collection respectively, to calculate the similarity included in the two objective function formulations defined in the next section. n denotes the number of sentences in the document collection.

$$sim(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}}, \quad i, j = 1, ..., n \tag{2}$$

2.2 Optimization

Coverage and diversity are formulated as two objective functions of the multi-objective optimization problem that are to be maximized. The first objective function, coverage (Eq. 3), ensures that the summary contains maximum amount of information. On the other hand, diversity (Eq. 4) reduces redundant information within the summary. So the maximization of diversity is used as a conflicting objective function. The rationale behind this is that there is a tendency for the summary to cover the original content more as the sentences are added to the summary, but at the same time, diversity of the content in the summary decreases. However, due to the limitation in the maximum number of sentences in the summary, we want to minimize redundant sentences being added to the summary. So the diversity is used as a countermeasure for the coverage. Multi-objective optimization algorithm ensures finding non-dominating solutions by considering these two conflicting objectives. Once we get the Pareto optimal solutions, we select the maximum coverage solution as our final summary.

$$f_{coverage}(\mathbf{X}) = G \cdot \sum_{i=1}^{k} \sum_{j=1}^{n} sim(c_i, s_j) \cdot x_j \tag{3}$$

$$f_{diversity}(\mathbf{X}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sim(s_i, s_j) \cdot x_i \cdot x_j \tag{4}$$

$\mathbf{X}$ is a binary vector where each element x_i is 1 if s_i is included in the summary, and is 0 otherwise. G represents global factor which measures similarity between the original document and the summary, and is calculated by comparing similarity between average weight of the sentences in the original document set and the average weight of the sentences in a candidate summary. We apply modifications in the initialization, crossover, and mutation part of the existing evolutionary multi-objective optimization algorithm to accommodate the 200 words limit constraint by adding or removing sentences from the summary.

3 EXPERIMENTS AND RESULTS

Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [1] is used to optimize the objective functions. The experiments are performed with the Document Understanding Conference (DUC) dataset of year 2002. DUC2002 dataset contains 59 topics with 5 to 10 documents included in each topic. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric is used to verify our experimental results against the gold standard summaries generated by human annotators. The ROUGE score measures how much the words in the human reference summaries overlap with the machine generated summaries in terms of n-gram.

Figure 2 shows the comparison of proposed MS-EMO with existing methods. The preliminary experiment on DUC2002 dataset shows that the proposed method successfully generates effective summaries is comparable to the state-of-the-art method as well as providing multiple Pareto optimal solutions on top of which we can build data analysis method to select the most similar summary to the human generated summary.

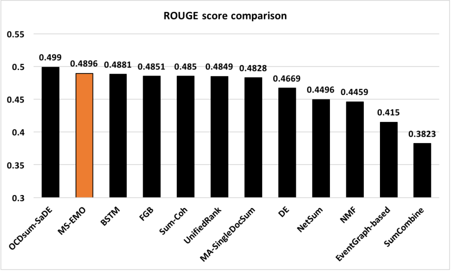

Figure 2: Comparison of existing methods[2] with MS-EMO.

4 CONCLUSION

An evolutionary multi-objective optimization based approach for multi-document summarization, namely MS-EMO, is proposed. The conflicting objective functions (coverage and diversity) drive the summary to cover maximum amount of information while ensuring maximum diversity between the summary sentences. Our initial experiment shows that the proposed method is efficient to generate automatic summaries using ROUGE evaluation metric.

ACKNOWLEDGMENT:

This work is funded by the ICT RD program of MSIP/IITP. [2016-0-00563,Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion]

REFERENCES

[1] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.

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

[3] Adam Kilgarriff and Christiane Fellbaum. 2000. WordNet: An Electronic Lexical Database. (2000).

[4] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.