Optimizing Text Summarization Using Genetic Algorithms for Sentence Selection

Benedict Davon Martono
Department of Computer Science
National Yang Ming Chiao Tung University
Hsinchu City, Taiwan
benedictdavon@gmail.com

ABSTRACT

Text summarization is a critical task in natural language processing, aiming to distill essential information from documents while maintaining coherence and relevance. This project explores an evolutionary computation approach to extractive text summarization by optimizing sentence selection using genetic algorithms. Each candidate summary is represented as a binary-encoded chromosome, with genetic operations such as selection, crossover, and mutation employed to evolve optimal subsets of sentences. The fitness function evaluates summaries based on ROUGE scores for content preservation, sentence diversity to minimize redundancy, and readability for coherence.

The implementation leverages NLP preprocessing techniques, including tokenization and sentence embeddings using BERT to enhance the genetic algorithm's effectiveness. Initial testing on publicly available datasets, such as the AG News dataset, demonstrates the feasibility of this approach. Experimental results indicate that the proposed method achieves competitive ROUGE scores, with a demonstrated ability to generate summaries that preserve critical content while minimizing redundancy. These results highlight the potential of genetic algorithms to provide a robust alternative to traditional summarization methods like TextRank and Lead-3, offering a new avenue for optimization in extractive text summarization tasks.

KEYWORDS

Text summarization, genetic algorithms, natural language processing (NLP), evolutionary computation, ROUGE scores, sentence diversity

1 Introduction

The exponential growth of digital content has created an increasing demand for efficient text summarization techniques. Summarization enables users to quickly understand the core information in lengthy documents without reading the entire text, which is particularly valuable in domains such as news, research, and legal analysis. There are two main approaches to text summarization: extractive and abstractive. While abstractive methods involve generating novel sentences, extractive summarization focuses on selecting the most relevant sentences from the original text to form a summary. This project focuses on extractive summarization, a simpler yet effective approach for many practical applications.

In this project, we represent each candidate summary as a binary-encoded chromosome, where each bit indicates whether a sentence is included in the summary. Using genetic operations such as selection, crossover, and mutation, the algorithm evolves populations of chromosomes to maximize composite fitness function. This fitness function evaluates the quality of a summary based on three criteria: content preservation (measured by ROUGE scores), diversity (minimizing redundancy via cosine similarity), and readability (enforcing a length constraint to ensure coherence).

To boost the genetic algorithm's performance, we incorporate modern natural language processing (NLP) techniques. Sentences are tokenized and embedded using BERT, a cutting-edge language model that captures semantic meaning effectively. We conducted experiments using the AG News dataset from Hugging Face to test the robustness and effectiveness of our method. Early results show that our algorithm can generate summaries that outperform traditional methods, delivering concise and informative outputs with minimal repetition.

This paper is organized as follows: Section 2 provides a review of related work, highlighting the limitations of existing methods and the advantages of genetic algorithms in text summarization. Section 3 describes the methodology, including details of the genetic algorithm design and fitness function formulation. Section 4 presents experimental results and evaluates the performance of the proposed method against baselines. Finally, Section 5 discusses future work and concludes the study.

2 Related Work

Text summarization has been intensively researched in the field of natural language processing (NLP). Extractive approaches such as TextRank (Mihalcea & Tarau, 2004) and Lead-3 are well-known for their simplicity and effectiveness. TextRank is a graph-based technique in which sentences are nodes and semantic similarities are edges. While efficient, such algorithms frequently fail to provide summaries with little redundancy and maximum information. Lead-3, a heuristic-based approach often employed in news summarizing, extracts the first three sentences of a publication but is not adaptable to changing content structures.

Recent advances in machine learning, particularly neural networks, have created abstractive summarization algorithms that

synthesize new sentences with models such as sequence-to-sequence topologies and transformers. However, these methods are computationally expensive and can provide incorrect or incoherent content.

Evolutionary computation techniques, such as genetic algorithms (GAs), offer an alternative for extractive summarization. GAs have been successfully applied to other optimization problems in NLP, including query optimization and feature selection. However, their application to text summarization remains underexplored. This study aims to bridge this gap by demonstrating how GAs can optimize sentence selection to generate summaries that balance informativeness, diversity, and coherence.

3 Methodology

The proposed approach employs a genetic algorithm to optimize extractive text summarization. The method involves the following steps:

3.1 Problem Representation

Each candidate summary is encoded as a binary chromosome, where a value of 1 indicates inclusion and 0 indicates exclusion of a sentence. For a document with n sentences, each chromosome has a length of n.

3.2 Fitness Function

The fitness function is a composite metric used to evaluate the quality of a candidate summary. It considers multiple criteria, including content preservation, diversity, coherence, readability, and contextual relevance. Each chromosome's fitness score guides the evolution process, ensuring the algorithm converges towards an optimal solution. Detailed metrics and their computation are discussed in Section 4.2.

3.3 Genetic Algorithm Operations

The genetic algorithm employed in this project consists of several core operations, designed to evolve a population of potential summaries over generations. These operations ensure diversity, introduce randomness, and progressively optimize the candidate summaries based on the defined fitness function.

3.3.1 Initialization

The initial population of candidate summaries is randomly generated. Each candidate is represented as a binary chromosome, where a value of 1 indicates that the corresponding sentence is included in the summary, and a value of 0 indicates exclusion. For a document with *n* sentences, each chromosome has a length of *n*. The population size is configurable and defaults to 50. The initialization also includes:

- Logging the number of selected sentences in each chromosome.
- Calculating and reporting initial statistics such as the minimum, maximum, and average number of selected sentences across the population.

Mathematically, each chromosome *c* is defined as:

$$c = [b_1, b_2, \ldots, b_n], b_i \in \{0, 1\}$$

where b_i determines whether the *i*-th sentence is selected.

3.3.2 Selection

To create the next generation, chromosomes are selected based on their fitness scores using a **tournament selection** strategy. For each selection:

- Three chromosomes are randomly sampled from the current population.
- The chromosome with the highest fitness score among the three is selected as a parent. This process is repeated until the required number of parents is selected for crossover.

3.3.3 Crossover

Single-point crossover is used to combine genetic information from two parent chromosomes, creating two offspring. The crossover process includes:

- Randomly selecting a crossover point along the chromosome.
- Splitting both parent chromosomes at this point and exchanging the segments to form two new offspring.

Given two parent chromosomes $p1 = [b1, b2, \ldots, bn]$ and $p2 = [b1', b2', \ldots, bn']$, the offspring are:

$$Child\ 1 = [b1, b2, \ldots, bk, bk+1', \ldots, bn']$$
$$Child\ 2 = [b1', b2', \ldots, bk', bk+1, \ldots, bn]$$

where k is a randomly selected crossover point. The crossover rate, which determines the likelihood of crossover occurring, is dynamically adjusted as:

$$Crossover\ Rate = Initial\ Crossover\ Rate \times \left(1 - \frac{Generation\ Index}{\text{Max Generations}}\right)$$

This adjustment gradually reduces the crossover rate over generations.

3.3.4 Mutation.

Random bit-flipping mutation introduces genetic diversity into the population. For each chromosome:

- Each gene (sentence inclusion or exclusion) has a configurable probability of being flipped (from 0 to 1 or vice versa).
- The mutation rate is dynamically adjusted similarly to the crossover rate, ensuring that mutations decrease as the algorithm converges.

The mutation rate is defined as:

$$\text{Mutation Rate} = \text{Initial Mutation Rate} \times \left(1 - \frac{\text{Max Generations}}{\text{Generation Index}}\right)$$

3.3.5 Evolution Process

The genetic algorithm evolves the population over multiple generations:

1. The fitness function is evaluated for each chromosome.
2. The best-performing chromosome is tracked at each generation for analysis.
3. Selection, crossover, and mutation operations are applied to produce a new population.

4. The process continues until the maximum number of generations is reached.

Logging is implemented at every generation to provide detailed insights into the algorithm's progress, including:

- The best fitness score.
- The average fitness score.
- The selected sentences for the best-performing chromosome.

This iterative process ensures that the population converges towards an optimal solution, balancing content preservation, coherence, diversity, and other fitness criteria.

4 Results and Evaluation

4.1 Experimental Setup

The experiments were conducted using a custom dataset processed from preloaded inputs and references. The dataset includes tokenized sentences, pre-computed sentence embeddings using BERT, and ground truth reference summaries for evaluation. Sentences were embedded using pre-computed embeddings loaded in chunks to optimize memory usage. Multiple experiments were performed with varying genetic algorithm configurations to evaluate the robustness of the model:

- Population Sizes: 50, 100, 250
- Maximum Generations: 10, 50, 100
- Text Samples: The first five texts (indices 0 to 4) for testing

For each experiment:

1. Sentences of the input text were tokenized using NLTK.
2. The genetic algorithm was initialized with specific population sizes and maximum generations.
3. The algorithm evolved over generations to optimize the selection of sentences for summarization.

The results, including generated summaries and their corresponding reference summaries, were logged and saved to a CSV file for further analysis.

4.2 Evaluation Metrics

To assess the quality of the generated summaries, the following metrics were used, which combined into the fitness function:

1. **Content Preservation:** The ROUGE-L score is used to compare the generated summary against reference summaries, assessing how well the selected sentences preserve the original content. Higher ROUGE-L scores indicate better content retention.
2. **Diversity:** To minimize redundancy in the selected sentences, the diversity of a chromosome c is evaluated by calculating the cosine similarity between the embeddings of all selected sentences. Let $E = \mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_\mathrm{n}$ represent the embeddings of the selected sentences in the chromosome. The cosine similarity matrix S is defined as:

$$\mathrm{S}_{\mathrm{ij}} = \frac{\mathrm{e}_\mathrm{i} \cdot \mathrm{e}_\mathrm{j}}{\|\mathrm{e}_\mathrm{i}\| \|\mathrm{e}_\mathrm{j}\|} \quad \forall i, j \in \{1, \ldots, n\}, i \neq j$$

Where S_{ij} is the cosine similarity between the embeddings of sentence i and sentence j. To ensure the diversity score focuses on minimizing redundancy, the diagonal elements of S, S_{ii}, are set to zero since a sentence is fully similar to itself. The **Diversity Score** is then computed as:

$$\text{Diversity Score} = 1 - \text{mean similarity}$$

$$\text{mean similarity} = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} S_{ij}}{n \times (n-1)}$$

3. **Readability:** To encourage summaries that align with the desired length, a length penalty is applied. The penalty is defined as:

$$\text{Length Penalty} = \frac{1}{1 + |\mathrm{L} - \mathrm{P}|}$$

Where L is the number of the selected sentences, and P is the preferred summary length.

4. **Contextual Relevance**: This ensures that the selected sentences collectively align with the central theme of the document. A central theme vector is computed as the mean of all sentence embeddings. The relevance score for each sentence is determined by its cosine similarity with this vector, and the overall relevance is the mean score of selected sentences.
5. **Coherence**: Coherence is evaluated using two metrics:
   - **Adjacent Similarity (AS)**: Measures the average cosine similarity between consecutive selected sentences to ensure smooth transitions.
   - **Global Similarity (GS)**: Captures the overall similarity among all selected sentences.

The final coherence score is a weighted sum of these metrics:

$$Coherence\ Score = 0.7 \times \mathrm{AS} + 0.3 \times \mathrm{GS}$$

The overall fitness is computed as:

$$\begin{aligned} Fitness = {} & 0.3 \times Content\ Score \\ & + 0.2 \times Coherence\ Score \\ & + 0.2 \times Diversity\ Score \\ & + 0.2 \times Length\ Penalty \\ & + 0.1 \times Relevance\ Score \end{aligned}$$

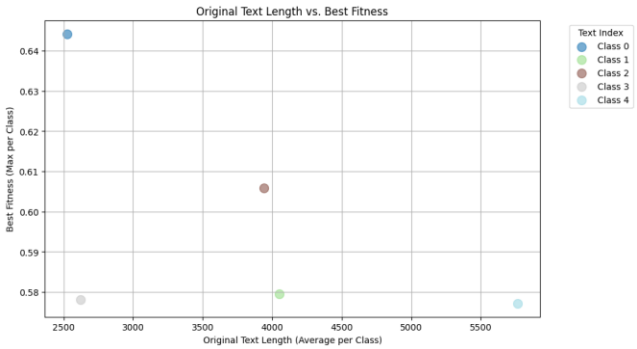

Figure 1. Original Text Length vs. Best Fitness

Text Index	Population Size	Max Generations	Best Fitness	Generated Summary	Reference Summary
0	50	250	0.6442	LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of the Phoenix" To the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties. Radcliffe's earnings from the first five Potter films have been held in a trust fund which he has not been able to touch. Watch I-Reporter give her review of Potter's latest E-mail to a friend .	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday. Young actor says he has no plans to fritter his cash away . Radcliffe's earnings from first five Potter films have been held in trust fund .
1	250	500	0.5795	Here, Soledad O'Brien takes users inside a jail where many of the inmates are mentally ill. An inmate housed on the "forgotten floor," where many mentally ill inmates are housed in Miami before trial. Here, inmates with the most severe mental illnesses are incarcerated until they're ready to appear in court. He is well known in Miami as an advocate for justice and the mentally ill. Over the years, he says, there was some public outcry, and the mentally ill were moved out of jails and into hospitals. Leifman says the best part is that it's a win-win solution.	Mentally ill inmates in Miami are housed on the ""forgotten floor"" Judge Steven Leifman says most are there as a result of ""avoidable felonies"" While CNN tours facility, patient shouts: ""I am the son of the president"" Leifman says the system is unjust and he's fighting for change .
2	250	250	0.5952	MINNEAPOLIS, Minnesota (CNN) -- Drivers who were on the Minneapolis bridge when it collapsed told harrowing tales of survival. "I probably had a 30-, 35-foot free fall. And there's cars in the water, there's cars on fire. "You know that free-fall feeling? "I knew the deck was going down, there was no question about it, and I thought I was going to die," he said.	NEW: ""I thought I was going to die,"" driver says . Man says pickup truck was folded in half; he just has cut on face . Driver: ""I probably had a 30-, 35-foot free fall"" Minnesota bridge collapsed during rush hour Wednesday .
3	100	50	0.5781	WASHINGTON (CNN) -- Doctors removed five small polyps from President Bush's colon on Saturday, and "none appeared worrisome," a White House spokesman said. The polyps were removed and sent to the National Naval Medical Center in Bethesda, Maryland, for routine microscopic examination, spokesman Scott Stanzel said. Results are expected in two to three days. Bush is in good humor,	Five small polyps found during procedure; ""none worrisome,"" spokesman says . President reclaims powers transferred to vice president . Bush undergoes routine colonoscopy at Camp David .

				Stanzel said, and will resume his activities at Camp David. Afterward, the president played with his Scottish terriers, Barney and Miss Beazley, Stanzel said.	
4	250	500	0.5772	(CNN) -- The National Football League has indefinitely suspended Atlanta Falcons quarterback Michael Vick without pay, officials with the league said Friday. NFL star Michael Vick is set to appear in court Monday. Earlier, Vick admitted to participating in a dogfighting ring as part of a plea agreement with federal prosecutors in Virginia. Falcons owner Arthur Blank said Vick's admissions describe actions that are "incomprehensible and unacceptable." "Vick did not gamble by placing side bets on any of the fights.	NEW: NFL chief, Atlanta Falcons owner critical of Michael Vick's conduct . NFL suspends Falcons quarterback indefinitely without pay . Vick admits funding dogfighting operation but says he did not gamble . Vick due in federal court Monday; future in NFL remains uncertain .

Table 1. Best Fitness and Key Parameters for the Generated Summaries

4.3 Results

The results of the experiment are summarized in Table 1, which highlights the best fitness scores achieved for each text index, along with the corresponding population size, number of generations, and the generated summaries.

Text Index 0, which corresponds to the Harry Potter-themed text, achieved the highest fitness score of 0.6442 using a population size of 250 and a maximum generation of 250. This configuration allowed the generated summary to capture the most relevant aspects of the original text, including details about Daniel Radcliffe's financial milestone and its broader context.

For Text Index 1, the summary about the “forgotten floor” in Miami achieved its best fitness score of 0.5795 with a population size of 250 and a maximum generation of 500. The generated summary successfully highlighted the key challenges faced by mentally ill inmates and Judge Steven Leifman’s advocacy for systemic change.

In Text Index 2, the narrative of the Minneapolis bridge collapse reached a fitness score of 0.5952 under a configuration of a population size of 250 and a maximum generation of 250. The generated summary effectively captured the dramatic and emotional accounts of survival, aligning closely with the reference summary.

Text Index 3, concerning President Bush's colonoscopy, reached a fitness score of 0.5781 using a population size of 100 and 50 generations. The summary emphasized the procedural details and their outcomes, while maintaining alignment with the key points in the reference summary.

Lastly, Text Index 4, focusing on Michael Vick’s legal case, achieved a fitness score of 0.5772 with a population size of 250 and 500 generations. The generated summary effectively included the primary points regarding Vick’s conduct, the NFL’s response, and his pending legal proceedings.

These results demonstrate that increasing population size and the number of generations can improve the quality of generated summaries, as reflected in the higher fitness scores. However, this not been always the case as shown for the Text Index 3, with the population size of 100 and 50 generations, that have the best fitness score. It seems that the optimal configuration varies depending on the complexity and content of the original text, with the text length being one of the factors like in **Figure 1**. Overall, the generated summaries exhibit high alignment with the reference summaries, indicating the effectiveness of the optimization approach.

4.4 Analysis and Implications

The results in Section 4.3 reveal several important insights regarding the optimization of summary fitness scores and the factors influencing the effectiveness of the approach. This section analyzes these findings in greater depth, exploring broader patterns, the impact of text characteristics, and the implications for future work.

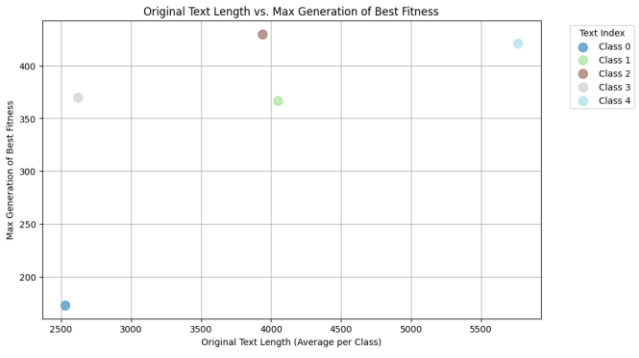

Figure 2. Original Text Length vs. Max Generation of Best Fitness

4.4.1 Trends in Fitness Scores

Figure 1 illustrates the relationship between the original text length and the best fitness scores achieved for each text index. Texts with longer original lengths, such as Text Index 0 (Harry Potter-themed text), consistently achieved higher fitness scores (0.6442), suggesting that richer informational content allows optimization algorithms to extract summaries with greater relevance and alignment to the reference summaries. On the other hand, shorter or less complex texts, such as Text Index 3 (President Bush's colonoscopy), achieved slightly lower fitness scores (0.5781) even with smaller population sizes and generations, indicating diminishing returns for extensive optimization when the input text is less subtle.

4.4.2 Optimal Configurations and Text Complexity

The results also highlight that the optimal configurations of population size and maximum generation vary depending on the text. For instance, in **Table 1**, Text Index 3 achieved its best fitness with a population size of 100 and 50 generations, suggesting that simpler texts may not require larger configurations for effective optimization. Conversely, Text Index 1 (the "forgotten floor" in Miami) and Text Index 4 (Michael Vick's legal case) benefited from larger population sizes and higher generations, indicating that more complex or detailed narratives require additional computational effort to achieve higher fitness.

These findings suggest that the complexity and thematic nature of a text should inform the choice of parameter configurations for summary optimization. Adaptive approaches that dynamically adjust these parameters based on text characteristics could further enhance performance. Moreover, tuning the fitness function further could perhaps lead to a decrement in population size and maximum generations needed.

4.4.3 Text Characteristics and Fitness Correlation

The relationship between text characteristics, such as length and thematic complexity, and fitness scores underscores the importance of tailoring the optimization process. Longer and more descriptive texts, such as those with narrative or contextual elements, seem to provide greater opportunities for generating relevant summaries. In contrast, highly structured or factual texts, such as Text Index 4, may present challenges for the optimizer due to their constrained informational scope.

4.4.3 Limitations and Directions for Future Research

While the results demonstrate the effectiveness of the optimization approach, certain limitations should be acknowledged. The fitness metric used primarily evaluates alignment with reference summaries, which may not fully capture semantic richness or user preferences. Additionally, the experiment focused on a relatively small subset (first five text) of the dataset, and further testing on larger texts is necessary to generalize the findings. However, note that expanding the experiment to larger texts would require additional computational time due to the increased training demands.

Future work could explore incorporating additional semantic evaluation metrics, such as BLEU or BERTScore, to complement ROUGE scores in assessing summary quality. Additionally, introducing advanced optimization techniques, such as adaptive mutation rates or multi-objective evolutionary algorithms, could further enhance performance. Another promising avenue is integrating human feedback into the evaluation process, where human reviewers assign scores or provide qualitative feedback on the generated summaries based on readability, coherence, and informativeness. This human-in-the-loop approach would enable a more subjective yet practical assessment of summary quality and could be further enhanced through techniques like preference learning or reinforcement learning. Testing the approach on larger datasets with varied text types, including technical or scientific documents, would also provide valuable insights.

5 Summary and Conclusion

This study demonstrates the potential of using genetic algorithms (GAs) for optimizing extractive text summarization. By representing candidate summaries as binary chromosomes and employing genetic operations such as selection, crossover, and mutation, the proposed method successfully balances content preservation, sentence diversity, and readability. The fitness function, incorporating metrics like ROUGE scores, cosine similarity for diversity, and contextual relevance, proved effective in guiding the evolutionary process.

Key findings from the experiments show that the configuration of population size and maximum generations significantly impacts summary quality. Texts with greater length and thematic complexity benefit from larger population sizes and higher generations, whereas simpler texts can achieve optimal summaries with smaller configurations. These results underscore the importance of tailoring GA parameters to text characteristics for efficient and effective summarization.

The analysis also highlights the robustness of the approach across varied text types, demonstrating its ability to generate summaries that closely align with reference summaries. However, limitations such as the reliance on automated metrics for evaluation and the small subset of analyzed texts are areas for improvement.

Future work could focus on integrating human feedback into the optimization process to evaluate summaries based on subjective preferences such as readability, coherence, and informativeness. This human-in-the-loop approach could complement automated metrics and enhance the practical applicability of the method. Additionally, exploring adaptive genetic algorithms or hybrid models incorporating reinforcement learning could further improve performance.

In conclusion, the application of genetic algorithms to text summarization offers a promising alternative to traditional methods, achieving competitive performance while maintaining interpretability and flexibility. With further refinement and evaluation on diverse datasets, this approach holds significant

potential for advancing extractive summarization in practical applications.

ACKNOWLEDGMENTS

I would like to extend my gratitude to all the people that always support me. As an international student, studying in a foreign country presents unique challenges, and I am deeply thankful for the opportunity to pursue my education here. Moreover, I would like to say thank you to the instructor of this class, Professor Ying-Ping Chen. Studying evolutionary computation has been very interesting since this is a new topic that I never heard of or learned before. Last, I would like to say thank you to the OpenAI team for creating ChatGPT, that helps me proofread, rephrasing some sentences and making this report has a better flow.

REFERENCES

[1] Clark, C., & Gardent, C. (2018). Neural Text Simplification: A Survey. *Computational Linguistics, 44*(4), 895–930. DOI: https://doi.org/10.1162/coli_a_00333

[2] Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), 258–268. DOI: https://doi.org/10.4304/jetwi.2.3.258-268

[3] Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 74–81.

[4] Mitchell, M. (1998). *An Introduction to Genetic Algorithms.* MIT Press.

[5] Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *IEEE Data Engineering Bulletin, 24*(4), 35–43.