Managing Genetic Algorithm Parameters to Improve SegGen, a Thematic Segmentation Algorithm

N.S. Saygili - GSU
T. Acarman - GSU
T. Amghar - LERIA
B. Levrat –LERIA, GSU

Presentation Outline

- Introduction
- Motivations
- SegGen Algorithm
- Tuning Parameters of SegGen
- Experimental Results
- Conclusion

Introduction

- This study proposed improvement approaches to SegGen is a method that it applies genetic algorithm for the text segmentation purposes.
- Due to the remarkable increasing of digital documents, it required to document processing on digital platform. Text segmentation is one of the important tools that are used in this area.

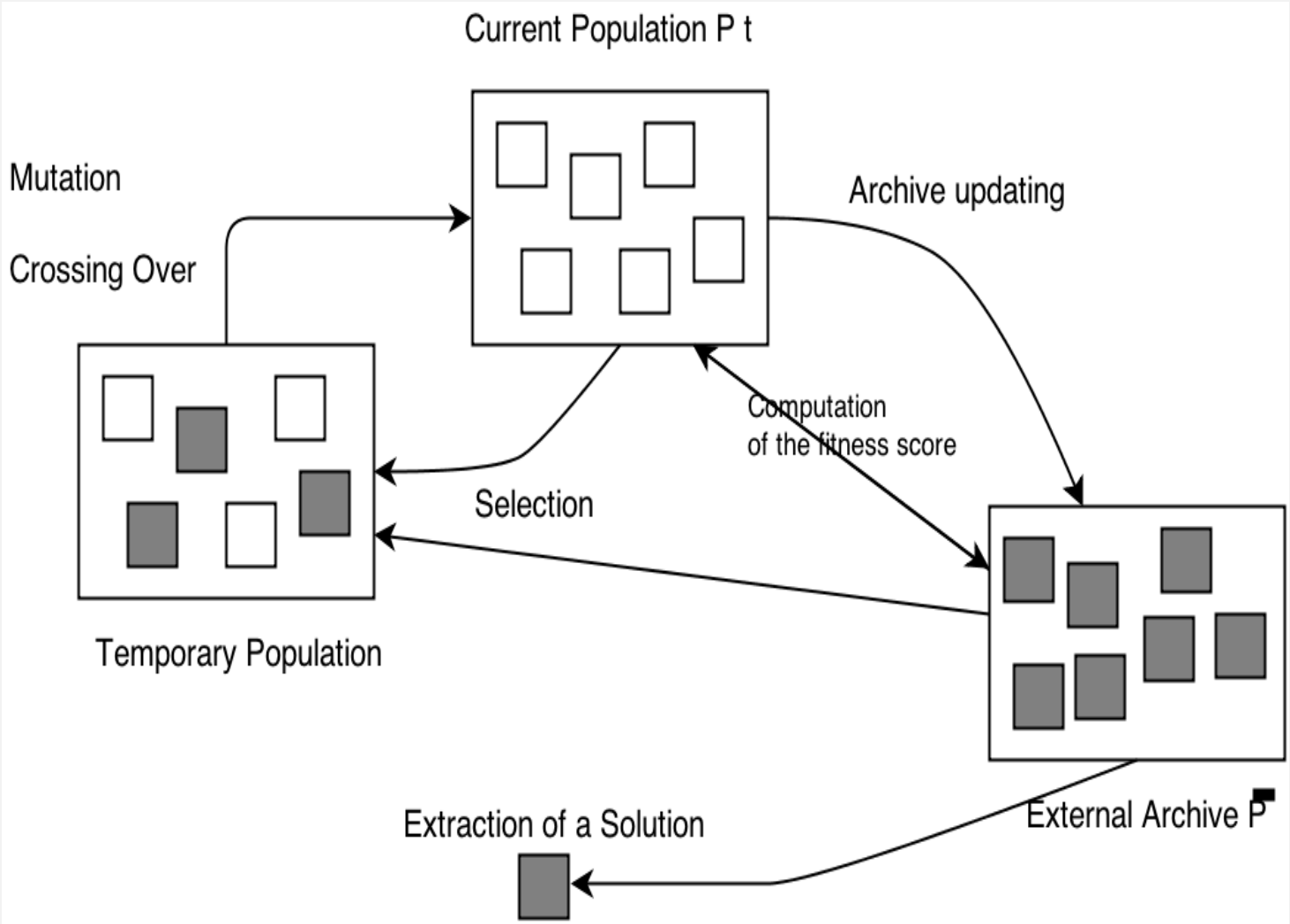
Motivations

- Similarities between sentences are examined locally nearby the potential segments and do not consider the whole potential segmentation.
- Most of the existing segmentation methods determine a sliding window for finding out dissimilarity measures in consecutive positions of the sliding window.
- However, the efficiency of such methods is very dependent of the dimension of the size of the sliding windows.

SegGen Algorithm

- SegGen algorithm permits to have a global view on all the potential segments to take a decision since all the boundaries between potential segments are set at the same time rendering.
- SegGen is a method that applies genetic algorithm for the text segmentation purposes.
- The main aim is to find out the subtopics, which create internal coherence and distinguished from other parts of the text.

General Flow of SegGen

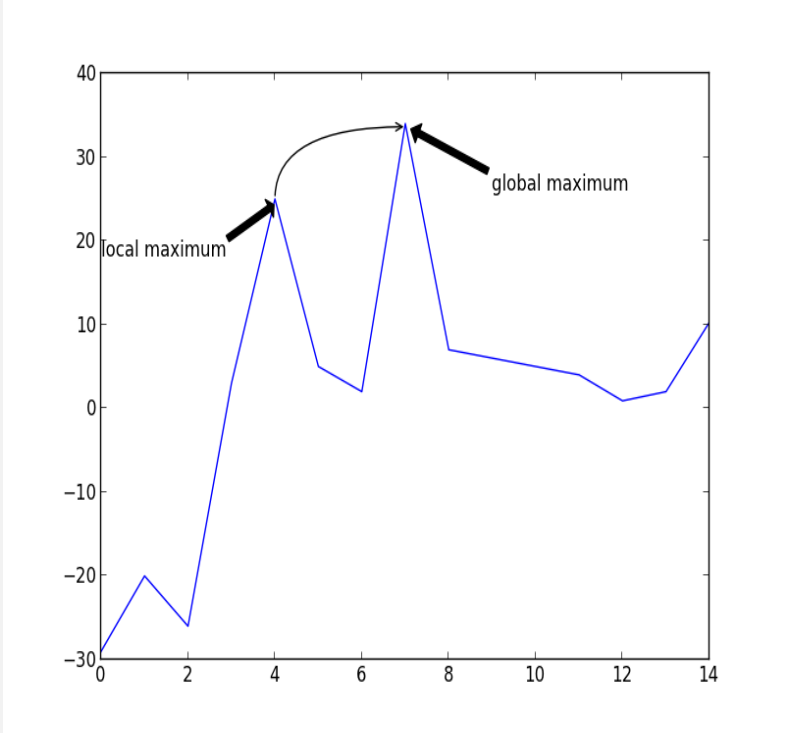


Proposed Improvement Approaches to SegGen

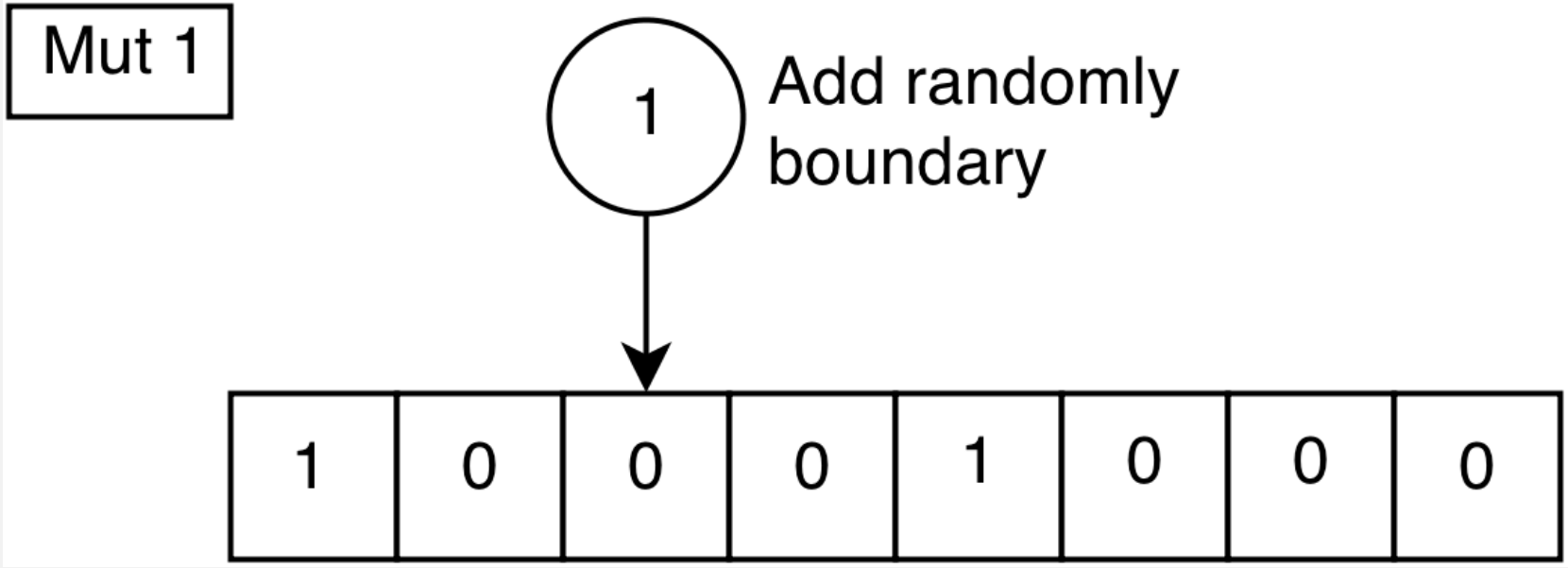
- 1) Genetic operators tuning , to set values to parameters and define new operators, which support intensification and reduce diversification factors in the search process.
- 2) Fitness function tuning, to weight sentences in the current segmentation depending on their distance to the boundaries of the segment.

Mutation Operator Tuning

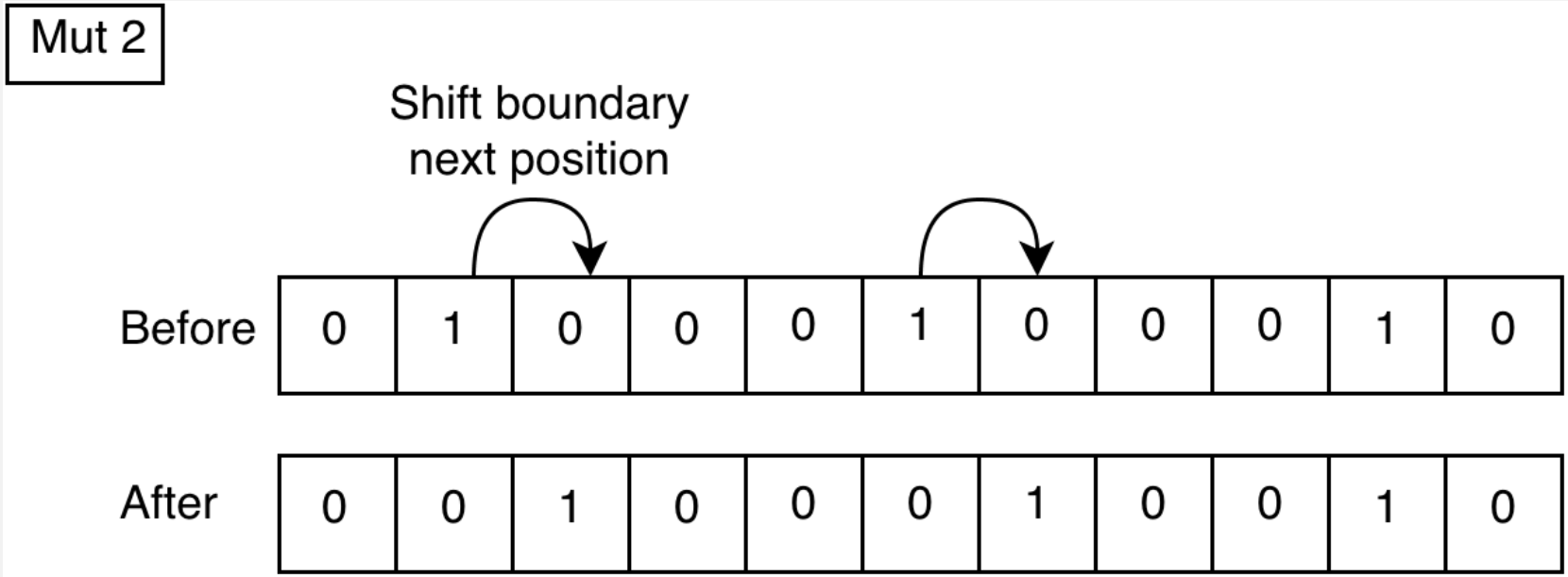
- The goal of the mutation operator is to diversify the search process.



Mutation Operator Tuning



Mutation Operator Tuning

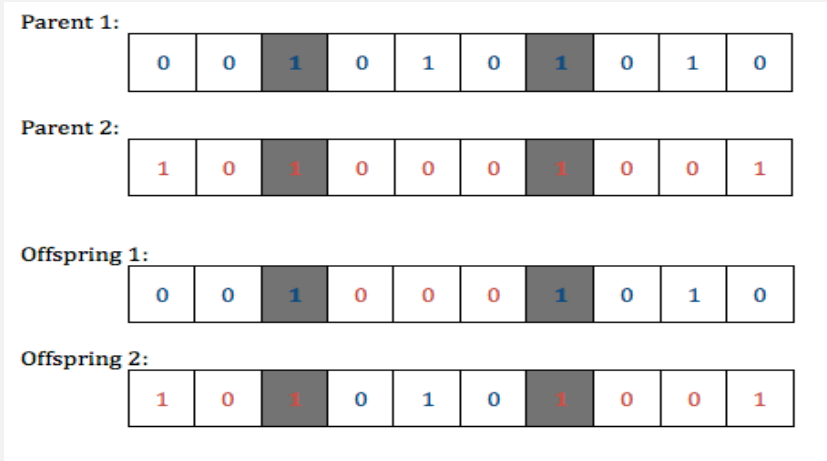


Mutation Probability Tuning

- We want to adjust mutation probability according to the quality of population.
- So we performed a trade off between mutation probability and average quality of population.

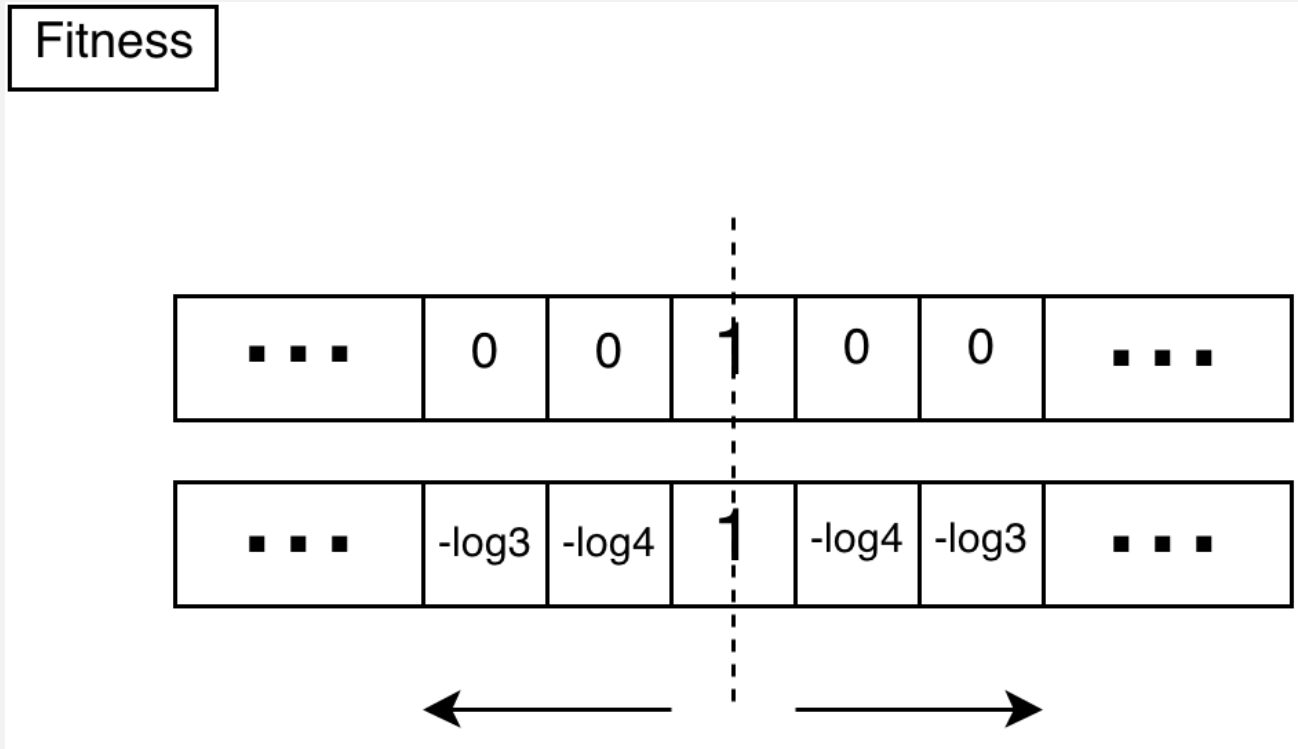
Crossover Operator Tuning

- A kind of proposed crossover that it uses two common boundary points of selected parents because the generated individuals have to keep existing boundaries on some part of the document to be defined.



Fitness Function Tuning

- We gave different importance value to sentences depends on their positions.



Experimental Results

- We concatenated sample articles which have various topics, selected from set of 350 documents.
- We created several test corpora are A(30,2), B(30,2), C(38,5), D(50,5), and E(55,7).
- We used a criterion *WindowDiff* as an evaluation metric.
- Windowdiff is a classical method that compares results by using a fixed-size sliding window.

Experimental Results

Windiff	Basic	M1	M2	C	M2C
A(30,2)	19.2	19.1	18.5	18.9	17.9
B(30,2)	20.1	21.3	20	20.5	18.7
C(38,5)	24.3	20.8	22.8	23.5	22
D(50,5)	25.6	24.5	22.8	24.7	24
E(55,7)	46.6	42.5	40.8	45.7	41.5

Experimental Results – cont'd

Windiff	Basic	Weighted	M2C-Weighted
A(30,2)	19.2	20.1	19.7
B(30,2)	20.1	22	21.5
C(38,5)	24.3	26.2	27
D(50,5)	25.6	25	26.3
E(55,7)	46.6	45.4	47.1

Experimental Results – cont'd

Windiff	Basic	Mixed	Weighted-Mixed
A(30,2)	19.2	16.5	22
B(30,2)	20.1	16.8	21.5
C(38,5)	24.3	22.2	24.2
D(50,5)	25.6	23.8	26
E(55,7)	46.6	41.4	42.5

Results

- The first results based on empirical values are promising.
- Tuned genetic operator versions of the algorithm results seem better than results of basic version of the algorithm.
- Especially, combination of all proposed tuning approaches is better than single versions.
- The fitness function tuning approach still needs some improvement such as tuning the weighted factor individual.

Conclusion

- In this thesis, we presented our improvement approaches to SegGen algorithm, which consists of tuning genetic operators and tuning fitness function.
- We have presented the first results of an ongoing work aiming at improving efficiency of SegGen.
- Even though, the parameters of the algorithm in first results rest upon empirical values, first results are promising perspectives.

Questions

- Thanks your attentions.
- Questions?