

Sentence Segmentation

*Report submitted in fulfillment of the requirements
for the Exploratory Project of*

Second Year B.Tech.

by

Rahul Chaturvedi

Under the guidance of

Dr Ravindranath Chowdary C



Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India
May 2019

Dedicated to

My parents, teachers

Declaration

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.
2. The work has not been submitted for any project.
3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.
4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi
Date: 03/05/2019

Rahul Chaturvedi
B.Tech. Student
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

*This is to certify that the work contained in this report entitled “**Sentence Segmentation**” being submitted by **Rahul Chaturvedi (Roll No. 17075047)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.*

Place: IIT (BHU) Varanasi
Date:

Dr Ravindranath Chowdary C
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Acknowledgments

I would like to express my sincere gratitude to Dr. Ravindranath Chowdary C for his guidance and constant supervision as well as for providing necessary information regarding the project and also for his support throughout the duration of the project. I would like to express my gratitude towards the people who have willingly helped me out with their abilities. Last but not the least, I would like to thank my parents for their due consideration to my busy schedule, providing me a constant motivation and support along with their unconditional love.

Place: IIT (BHU) Varanasi

Date:

Rahul Chaturvedi

Abstract

Sentence Segmentation, also known as Sentence boundary detection and Sentence breaking, is the problem in Natural Language Processing of dividing a string of written language into its component sentences.

In English and some other languages, using punctuation, particularly the full stop/period character is a reasonable approximation. However, even in English this problem is not trivial due to the use of the full stop character for abbreviations, which may or may not also terminate a sentence. For example, Dr. is not its own sentence in "Dr. Chowdary is very strict in terms of attendance."

Other languages like Chinese and Japanese, do not contain punctuation characters. Thus these languages have different rules for segmentation. This project, however, is focused on segmentation of English language only.

Contents

List of Figures	ix
List of Tables	ix
List of Symbols	x
1 Introduction	1
1.1 Overview	1
1.2 Motivation of the Research Work	1
2 Strategy and Algorithm Used	3
3 Algorithm in action	5
4 Conclusions and Discussion	9
Bibliography	11

List of Figures

3.1	Input page(Index)	5
3.2	My Output	6
3.3	NLTK Library Output	6
3.4	Input(Via File)	7
3.5	Output(1)	8
3.6	Output(2)	8

List of Tables

4.1	Comparison w.r.t NLTK library	9
-----	---	---

List of Symbols

Symbol	Description
Ψ	Field of Interest (FoI)
Ω	Boundary region of a FoI
b	Width of a region Ω
l	Length of a region Ω
ξ_i^k	Redundancy degree of a type i sensor for k -coverage of a FoI

Chapter 1

Introduction

1.1 Overview

The sentence is a fundamental and relatively well-understood unit in computational and theoretical linguistics. Natural language processing tasks often require the input(strings) to be divided into individual sentences; however, sentence breaking can be a difficult task due to the potential ambiguity of punctuations. In the English language, a period may indicate the end of a sentence but also may denote an ellipsis(...), a decimal point, an URL address or an abbreviation. Exclamation marks and Question marks can be similarly ambiguous due to their usage at other places.

1.2 Motivation of the Research Work

Sentence segmentation is an essential part of any Natural language processing system since the sentences and words identified at the stage are the fundamental units passed to further processing stages such as part-of-speech taggers, parsers, morphological analysers and information retrieval systems. In this project, I present you an approach for Sentence breaking of English language with high accuracy in comparison to other libraries which are already used for Sentence Boundary Detection such as OpenNLP,

Stanford NLP, NLTK. Here are some languages that do not use the same sentence ending punctuation (. ? !) as English:

- Amharic full stop, question mark
- Arabic question mark
- Armenian full stop
- Burmese full stop
- Chinese full stop, question mark, exclamation point
- Greek question mark
- Hindi full stop
- Japanese full stop, question mark, exclamation point
- Persian Urdu question mark

Chapter 2

Strategy and Algorithm Used

In the chapter,

I used Regular Expression(RegEx) module of python for the implementation since RegEx can be used to check and replace the pattern if a text contains the specified search pattern. I searched for all the abbreviations, decimal numbers, etc. and replaced them with /neos/ tag.

At first, I approached sentence boundary detection by first determining possible abbreviations. Quantitatively, abbreviations are a significant source of ambiguities in sentence boundary detection since they often constitute more than 40% of the potential candidates for sentence boundaries in running text. It is our underlying assumption that abbreviations are collocations of the truncated word and the following period, and hence that methods for the detection of collocations can be successfully applied to abbreviation detection.

Then, Secondly, I approached by determining possible decimal numbers and ellipsis. This step was the easiest of the lot since there is an easy pattern for both of them. But there were still some complications remaining in the model as it cant detect a period as the sentence boundary, written after a year.

Then, lastly, I approached for the edge cases which were found during the testing

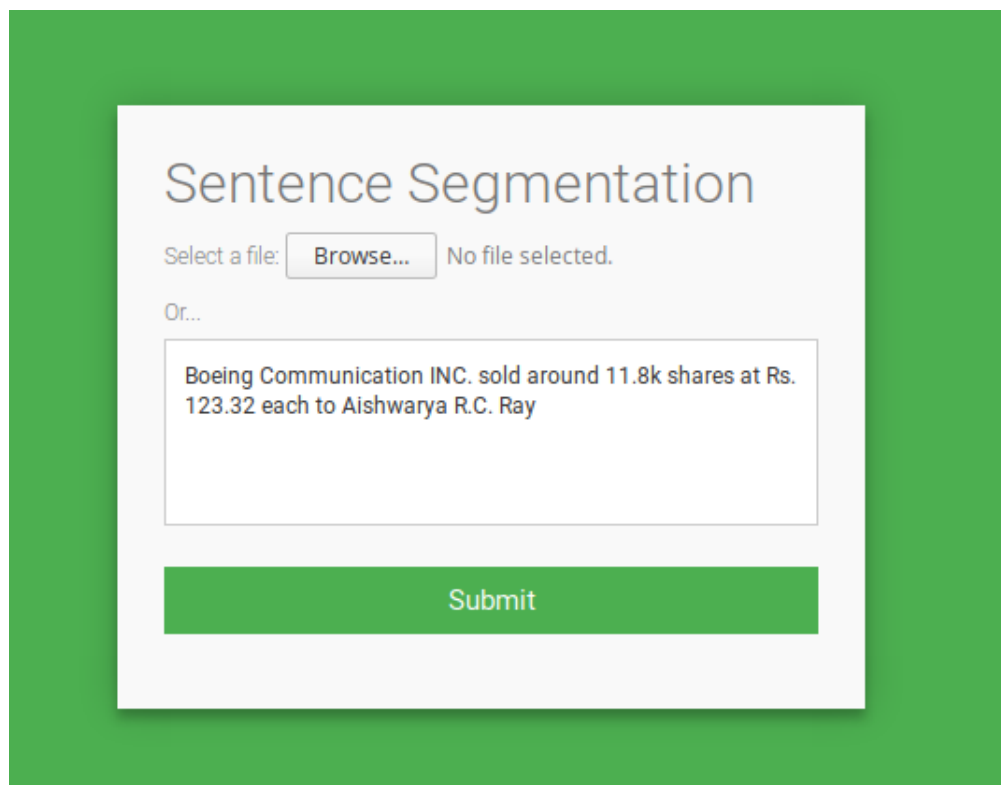
of the accuracy of my program in comparison to that of NLTK one. Some of the cases were URL detection, Sentence boundary detection inside quotes and brackets, which were quite complex and ambiguous.

Then, After writing my main code, I focused mainly on making an interface for my application, which takes a sentence/group of sentence and returns an output to the user in the GUI format. The interface was made with the help of Django, HTML and CSS.

Chapter 3

Algorithm in action

Here is an example where the input is made via the input box where my code is performing better than that of NLTK one.



Sentence Segmentation

Select a file: No file selected.

Or...

Boeing Communication INC. sold around 11.8k shares at Rs. 123.32 each to Aishwarya R.C. Ray

Figure 3.1 Input page(Index)

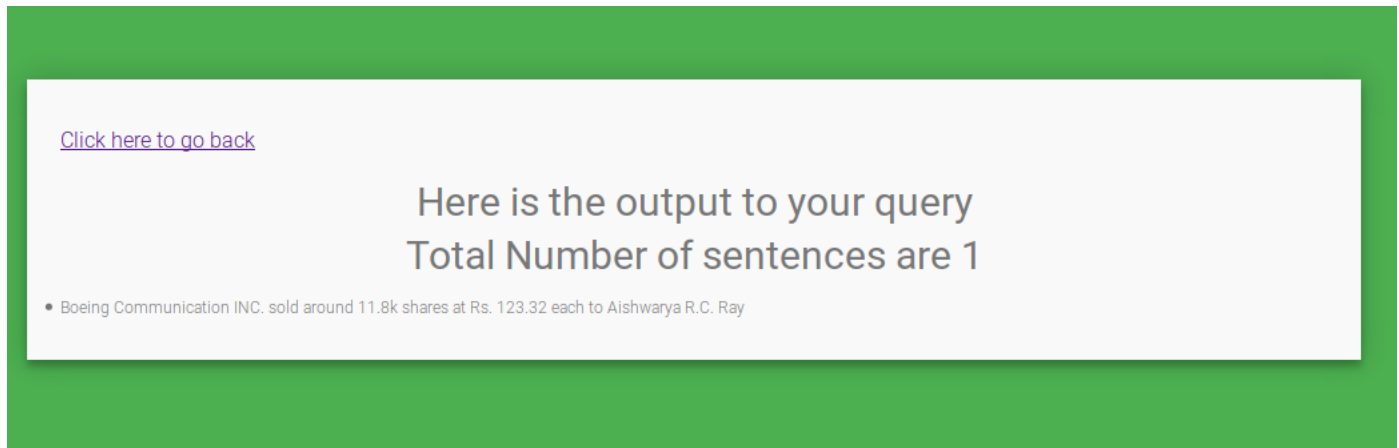


Figure 3.2 My Output

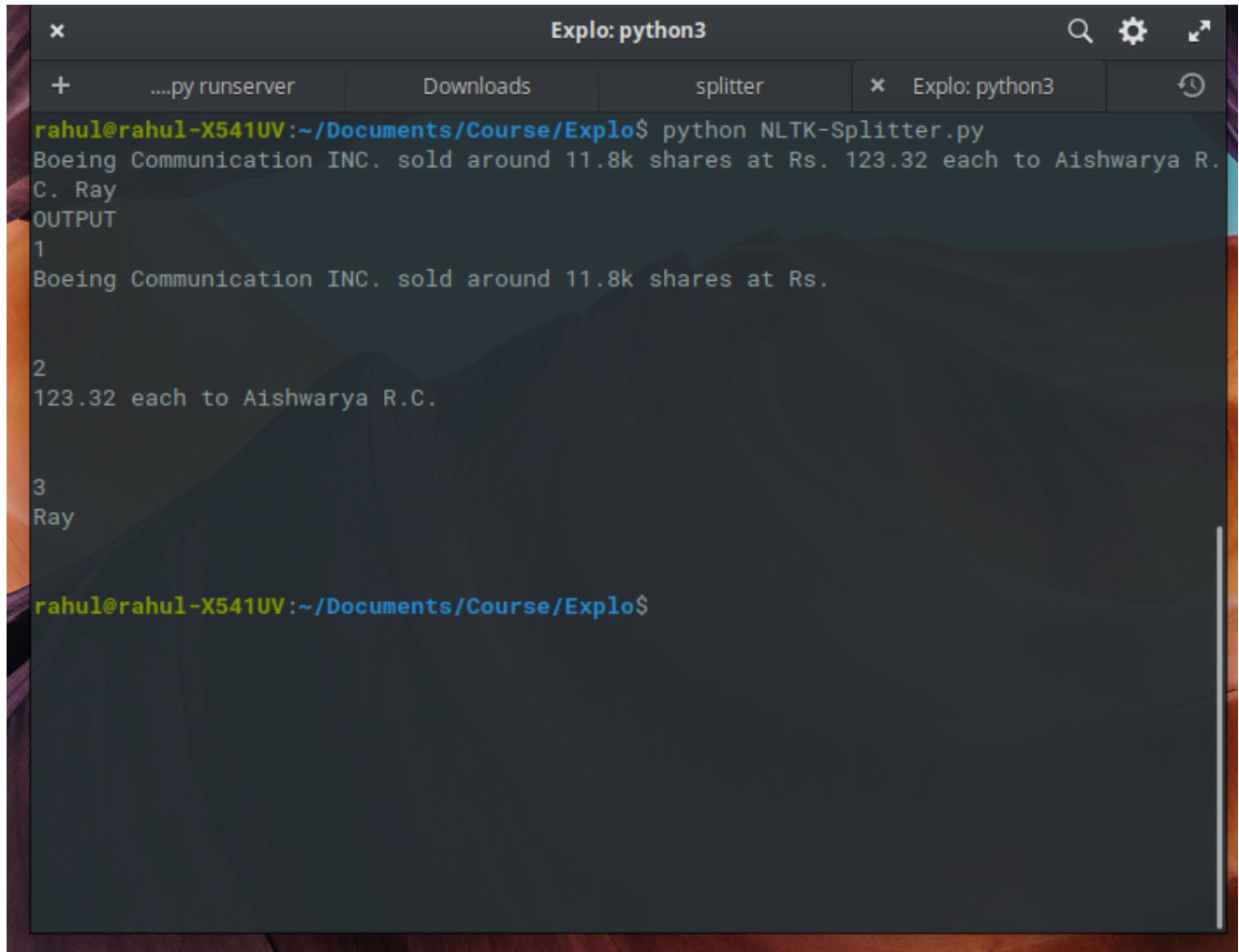


Figure 3.3 NLTK Library Output

Here is another example when the user uploads a file instead of writing manually in the box.

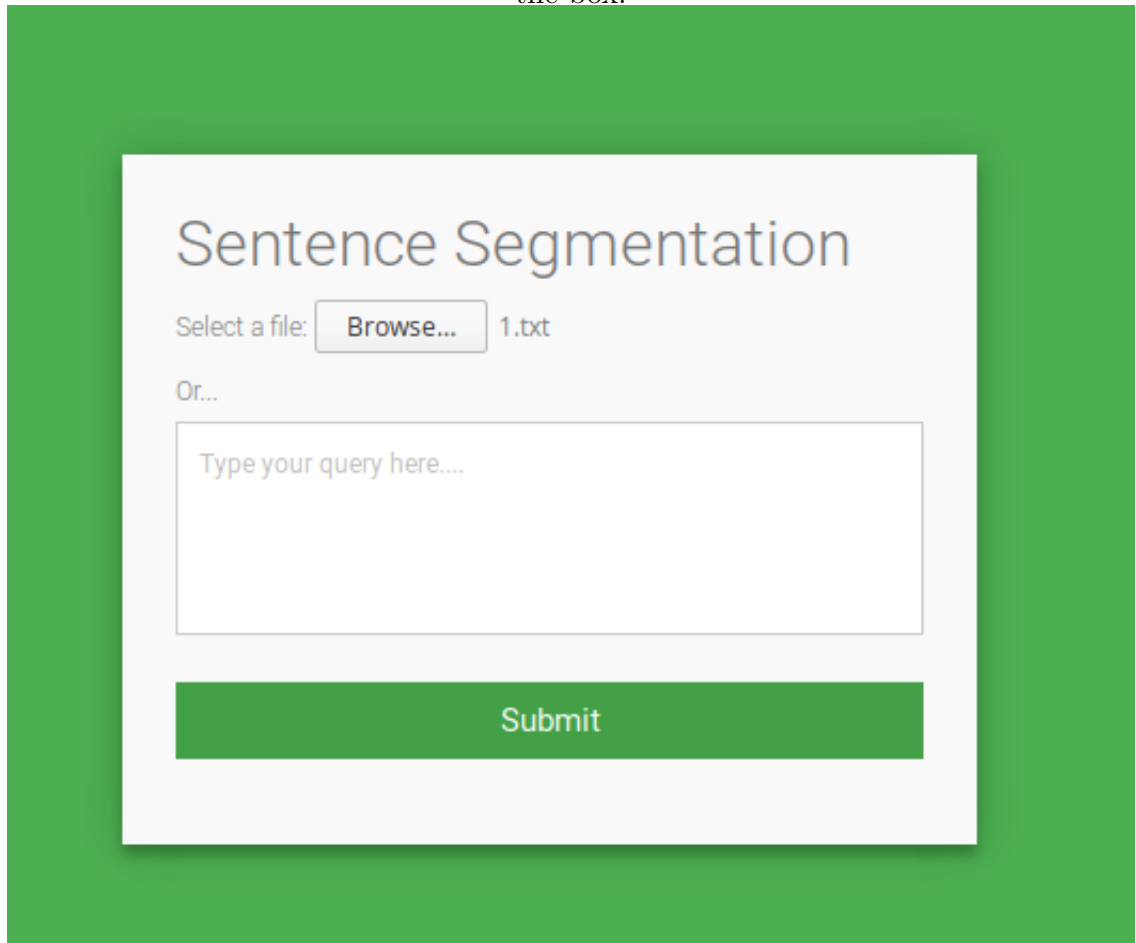
The image shows a web application interface for "Sentence Segmentation". It features a light gray rectangular box centered on a solid green background. Inside the box, the title "Sentence Segmentation" is displayed in a large, dark gray font. Below the title, there are two input options. The first option is "Select a file:", followed by a "Browse..." button and the text "1.txt". Below this, the word "Or..." is displayed. The second option is a large text input box with the placeholder text "Type your query here....". At the bottom of the gray box is a prominent green "Submit" button.

Figure 3.4 Input(Via File)

[Click here to go back](#)

Here is the output to your query Total Number of sentences are 19

- On the same ground where 10 days ago Muttiah Muralitharan raised expectations that he might take all 10 England wickets in the third Test, Charlie Shreck produced an equally destructive bowling display yesterday to lift Nottinghamshire off the foot of the First Division.
- All 10 was never a possibility for the lanky fast-medium swing bowler, because Middlesex's opening batsman Ben Hutton was absent with shingles.
- But he did take the first eight before the last man, Chris Silverwood, lofted Graeme Swann's off-spin to long-on to complete an ignominious three-day defeat.
- Middlesex crumbled to 49 all out in only 27.4 overs, the lowest championship total since the advent of two divisions in 2000.
- They now lie only one point ahead of Nottinghamshire - Yorkshire are now bottom - and, on this evidence, even allowing for the England one-day call-ups for Jamie Dalrymple and Ed Joyce, relegation cannot be discounted.
- Only Nick Compton, with 11 from 56 balls, committed himself to bloody-minded resistance as the rest succumbed timidly to Shreck's outswingers.
- His eight for 31 was the best return by a Nottinghamshire bowler since Ken Smales did take all 10 in Stroud against Gloucestershire in 1956.
- To outdo the likes of Richard Hadlee was quite a feat for a bowler who now has 40 wickets in six championship matches since his recovery from a back stress fracture and a useful winter in New Zealand.
- Part of his treatment for a curved spine was an order to stand up straight.
- Shreck saved his comments until the final whistle of England v Trinidad & Tobago; Middlesex were also comfortably showered and changed in time for the second half.
- "The ball swung from ball one," he said.
- "We had joked at tea that we might have it all over in time for the football.
- Last year was disappointing for me.

Figure 3.5 Output(1)

- His eight for 31 was the best return by a Nottinghamshire bowler since Ken Smales did take all 10 in Stroud against Gloucestershire in 1956.
- To outdo the likes of Richard Hadlee was quite a feat for a bowler who now has 40 wickets in six championship matches since his recovery from a back stress fracture and a useful winter in New Zealand.
- Part of his treatment for a curved spine was an order to stand up straight.
- Shreck saved his comments until the final whistle of England v Trinidad & Tobago; Middlesex were also comfortably showered and changed in time for the second half.
- "The ball swung from ball one," he said.
- "We had joked at tea that we might have it all over in time for the football.
- Last year was disappointing for me.
- It felt like another team winning the title."
- Two months into the season and belatedly Nottinghamshire, the defending champions, have stirred.
- David Hussey's 2005 championship batting average of 68 had played a substantial part in a first title for 18 years.
- His form had been dismal until his 107 yesterday, encompassing 239 balls and 5½ hours, his eighth-wicket stand of 117 in 21 overs with Swann ensuring a first-innings lead of 82.
- Swann's flaying of Johan Louw caused the South African fast bowler to demolish his sawdust pile and his mood darkened when he dropped a return catch when Swann was on 46.
- Hussey was last out, giving Silverwood his fifth wicket.

Figure 3.6 Output(2)

Chapter 4

Conclusions and Discussion

In this project, I examined total of 53 documents from various sources(in which some of the data was of unstructured manner) and calculated accuracy with respect to NLTK library one.

Percentage Range	Number of Documents
100	26
90-99	23
80-90	3
≤ 80	1

Table 4.1 Comparison w.r.t NLTK library

Many in the translation industry seem to think that sentence segmentation is a solved problem; however, I tend to disagree. While there are tools that can obtain high accuracy in specific languages and specific domains, there still does not exist a free, open-source tool that can handle many languages as well as handle ill-formatted content across any incoming domain. Although segmentation is not a very interesting problem, it is very important and the base of many other NLP functions and tasks (e.g. machine translation, bitext alignment, named entity extraction, part-of-speech tagging, summarization, classification, etc.). As segmentation is often the first step needed to perform these NLP tasks, poor accuracy in segmentation can lead to poor

end results.

In my opinion sentence segmentation is not yet a solved problem, but hopefully many NLP programmers can help to continue pushing the community forward by working to improve segmentation accuracy across many languages. It is certain that more work still needs to be done to further improve accuracy rates across all languages and to add support for even more languages.

Bibliography