Automatic Text Summarizer

Rohit Shukla (B11028)
Saurabh Jain (B11033)
Shubham Ajmera (B11035)

Problem Statement

- Automatic Text summarization by using Natural Language Processing and Machine learning algorithms.
- The summary generated will be an extract based summary rather than being a abstract one.
- It will be good enough for the reader to get the main idea of the document.

Progress till now…..

- The approach of "Extracting Sentence Segments" for text summarization has been implemented.
- Apache OpenNLP which is a Java machine learning toolkit for natural language processing(NLP) is used.
- Research paper by Wesley T. Chuang and Jihoon Yang of UCLA, Los Angeles has been referenced to implement the whole algorithm.

Apache OpenNLP

- A machine learning based toolkit for the processing of natural language text.

- Includes a tokenizer, a sentence detector, part-of-speech tagger, a name finder, a chunker and a parser.

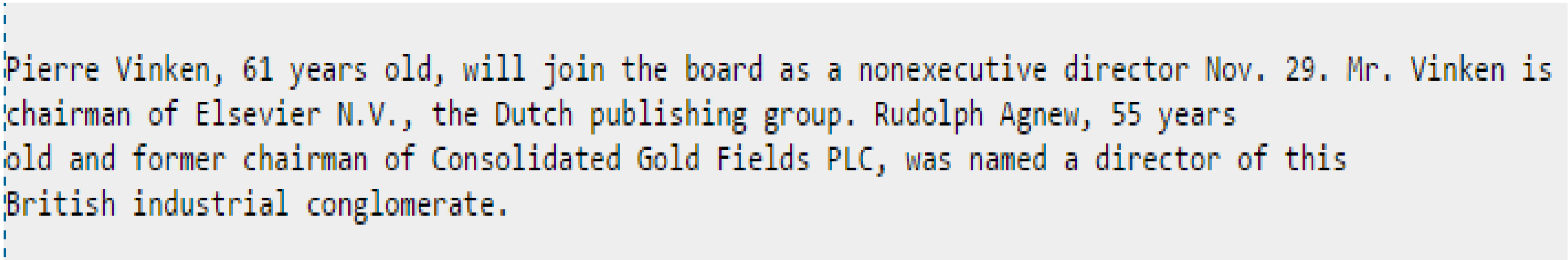
Sentence Detector

- Detects that a punctuation character marks the end of a sentence or not.

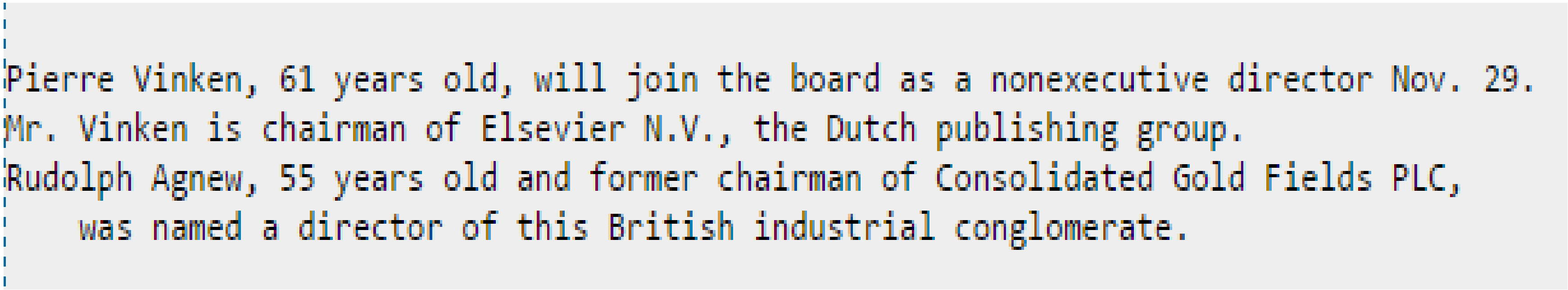
- It cannot identify sentence boundaries based on the contents of the sentences.

Contd…

- Example: The paragraph in the first image is broken down into sentences by the detector.



```
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vinken is
chairman of Elsevier N.V., the Dutch publishing group. Rudolph Agnew, 55 years
old and former chairman of Consolidated Gold Fields PLC, was named a director of this
British industrial conglomerate.
```



```
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC,
    was named a director of this British industrial conglomerate.
```

Sentence Segmentation

- Separate out units that convey independent meanings.
- A sentence is segmented by a cue phrase. Cue phrases are terms which separate a sentence into two segments.
- Cue phrases include terms such as : because, but, if, however etc. these terms define the rhetorical relations between the segments.

Contd….

- Segments generated are split into two clauses: the main segment nucleus and subordinate called satellite.

- Nucleus is generally considered as more important segment as compared to satellite.

- Segments are bounded by the bracket “[]” with an integer numbering their sequence . Words enclosed by “}” are called comma parenthesis which are considered as additional information inside a segment and will be thrown out in final summary.

Document

“ This invention relates in general to database management systems performed by computers, and in particular, to a method and apparatus for accessing a relational database over the Internet using macro language files. With the fast growing popularity of the Internet and the World Wide Web (also known as "WWW " or the " Web") there is also a fast growing demand for Web access to databases. However, it is especially difficult to use relational database management system(RDBMS) software with the Web. One of the problems with using RDBMS software on the Web is the lack of correspondence between the protocols used to communicate in the Web with the protocols used to communicate with RDBMS software. For example, the Web operates using the HyperText Transfer Protocol(HTTP) and the HyperText Markup Language(HTML).This protocol and language results in the communication and display of graphicalinformation that incorporates hyperlinks. Hyperlinks are network addresses that are embedded in a word, phrase, icon or picture that are activated .”

Segments

This invention relates in general to database management systems performed by computers, and in particular, to a method and apparatus for accessing a relational database over the Internet using macro language files. 1]

[With the fast growing popularity of the Internet and the World Wide Web { (also known as " WWW " or the " Web ") },2]

[there is also a fast growing demand for Web access to databases. 3]

[However, it is especially difficult to use relational database management system { (RDBMS) } software with the Web. 4]

[One of the problems with using RDBMS software on the Web is the lack of correspondence between the protocols used to communicate in the Web with the protocols used to communicate with RDBMS software.5]

[For example, the Web operates using the Hypertext Transfer Protocol { (HTTP) } and the Hypertext Mark up Language { (HTML) }. 6]

[This protocol and language results in the communication and display of graphical information that incorporates hyperlinks. 7]

Text features

- Sentence Position
- Positive keyword in sentence
- Sentence centrality
- Title Words
- Data such as Name, Places, Date/Time etc.
- Length

References

- opennlp.apache.org/documentation
- Extracting Sentence Segments for Text Summarization: A Machine Learning Approach

  by- Wesley .T . Chuang and Jihoon Yang
- Automatic Text Summarization by-Mohamed Abdel Fattah and Fuji Ren
- International Journal of Computer Science Volume 3 Number 1