Review Paper on Text Summarization using Natural Language Processing

Tanuj S Kulkarni
Dept. Of Computer Science Engineering
KLS Gogte Institute of Tech.
Belgaum, India
tanuj21199@gmail.com

MD. Shoeb Meti
Dept. Of Computer Science Engineering
KLS Gogte Institute of Tech.
Belgaum, India
mdshoebmeti@gmail.com

***Abstract*—Natural Language Processing (NLP) is a branch of computer science and artificial intelligence which deals with the understanding of human languages by a computer. NLP provides a way of analyzing natural languages by computerized means. NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. Time is very precious and there is a large amount of data being produced daily. Being updated to the new things in a specific field is a challenge, and it takes a lot of time to go through that much amount of data. NLP can be used to get the summary of that large data. This paper covers the evolution of Text Summarization using NLP.**

Keywords—NLP, machine learning, artificial intelligence

I. Introduction

Natural Language could be any language that humans speak in order to express knowledge and emotions and to convey responses to other people and surroundings. Natural languages are usually learnt from observing our surrounding in early childhood. Technology has not yet advanced to the point where these languages in all of their unprocessed forms can be understood by the computers. Natural Language Processing (NLP) is an area of research and application that explores how these languages can be manipulated and analysed by the computers to understand them. The field of NLP is very deep and diverse. NLP provides a range of computational techniques used to extract grammatical structure and meaning from naturally occurring text or speech, at one or more levels of linguistic analysis in order to perform useful tasks. The term NLP is normally used to describe the function of software or hardware components in a computer system which analyse or synthesize spoken or written languages.

One of the most popular application of NLP is Text Summarization. Text Summarization is the process of extracting the key information from a large amount of text data. With the summary generated the whole data can be understood easily. NLP is used to recognize the key components in the data by various methods. These methods evolved through the years and gave a wide description on the concepts that can be used to process a natural language. Text summarization is widely used in news articles generation, automated document review, mail client, report generation etc.

This paper will give a brief understanding behind the concepts used in each stage in evolution of text summarization.

II. LITERATURE REVIEW

The research work in natural language processing has progressed rapidly in the recent years. The natural language processing is a computerized approach to analyse text or speech. It has been a very active area of research and development lately. The literature describes the evolution of methods used for text summarization.

A. Positional method

First and last sentence of a paragraph are topic sentences (85%vs 7%).

B. Luhn's method

Frequency of contents terms in a sentence tell about the importance of the sentence.

C. Edmundson's method

This uses Bonus words (important words), Stigma words (negative words), Null words (irrelevant words) to find out the important sentences.

D. FRUMP scripts

This is a template filling approach based on UPI new stories

E. Classification

Uses Naïve Bayes classifier and assuming statistical independence of the features to compute the summary.

F. Mead

Centroid based method of computing summary.

G. LexRank

Graph based method that represents sentences as nodes in the graph and connecting nodes based on similarity matrix to compute the summary.

III. LEVELS OF NLP

By the means of 'levels of language' approach we will demonstrate what actually happens in NLP system.

A. Phonology:

This level deals with the pronunciation of speech sounds.
Rules used in Phonological analysis.

1) Phonoetic rules: It is used for sounds within words.

2) Phonemic rules: It is used for caritaions of pronunciation when words are spoken.

3) Prosodic rules: It is used to check for fluctuation in stress and intonation across a sentence.

B. Morphology:

The first stage of analyzing input after receiving it is Morphology. It determines the grammatical status of the words by breaking them down into components if possible.

Morphology is mainly useful for identifying the parts of speech in a sentence and words that interact together.
Morphology is a systematic description of words in a natural language. It describes a set of relations between words' surface forms and lexical forms.
The information gathered at the morphological stage prepares the data for the syntactical stage which looks more directly at the target language's grammatical structure.

1) Syntax: It is a task to determine role of each word in a sentence depending on the grammar defined by the target language.

C. *Semantics:*

It determines the actions a sentence is describing and has details provided by adjective, adverbs, propositions.

D. *Pragmatics:*

It is the "the analysis of the real meaning of an utterance in a human language, by disambiguating and contextualizing the utterance". This is done by identifying and removing ambiguities from the system using disambiguation techniques.

IV. METHODS AND APPROACHES

A. *Positional method*

- Introduced by P. Baxendale in 1958.
- Man-made index for technical literature.
- First and last sentence of a paragraph are topic sentences (85% vs 7%).

B. *Luhn's method*

- Introduced by H.P.Luhn in 1958.
- Computation of summary based on Frequency of content terms.
- Data pre-processing was introduced in this method for the first time. Methods like removal of stop words, Steaming.
- The method depends on selecting the sentences with highest concentration of salient content terms.

C. *Edmundson's method*

- Introduced by H. P. Edmundson in 1968.
- This method uses the Position of the sentence in a paragraph (P), Word frequency (F), Cue words(C) and Document Structure (S). The Cue words can be classified into Bonus words (important words), Stigma words (negative words), Null words (irrelevant words).
- The Score for each sentence was Linear combination of these 4 features: $a_1P + a_2F + a_3C + a_4S$

D. *Frump scripts*

- Introduced by G. deJong, 1979.
- This summarizer was a Fast Reading Understanding and Memory Program.
- This was a Knowledge based summarization method.
- It was a template filling approach based on UPI news stories. It used 50 sketchy scripts that contain important events that are expected to occur in a specific situation, Summarizer looks for instances of salient events, filling in as many as possible.

E. *Classification*

- Introduced by Kupiec et al. in 1995.
- It is the first trainable Document Summarizer. Training set: original documents and manually created extracts.
- This uses the Naïve Bayes classifier for text ranking defined by:

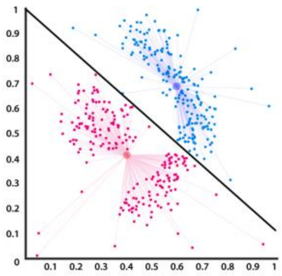
$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

- Performance of this method:

  For smaller summarises-74% improvement over lead summaries.

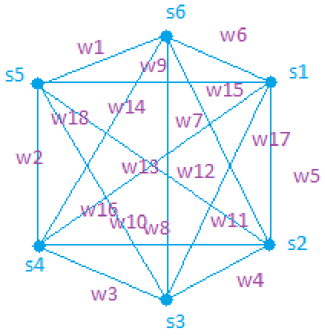
F. *Mead*

- This method is centroid-based. It assigns the sentences in a classifier and calculates the centroid for it. This centroid is used for calculating the sentence importance based on the distances from the centroid.



G. *LexRank*

- Lexical centrality
- It creates a similarity matrix. It represents sentences as nodes in the graph. Connecting nodes based on similarity matrix.
- Preferred for multiple document analysis.



V. CONCLUSION

While NLP is a recent area of research and application as compared to other technology approaches there have been a sufficient success to date. That suggest that NLP based information access technologies will continue to be a major area of research and development in the future. The art of NLP techniques applied to speech technologies, document analysis specifically for summary generation. We understand the concepts that were used in various methods of text summarization like, First and last sentence of a paragraph are topic sentences, centroid based concept etc.

This gives a brief idea of the important concepts in natural languages that can be used for text summarization. We can conclude by saying that Lex Rank Summarization technique is more suitable for dealing with data from multiple documents and also we can say that depending on the format of description in the data, different methods can used to get the summary more accurately.

VI. REFERENCES

[1] Wohleb, R. "Natural Language Processing: Understanding Its Future," PC/AI, November/December, 2001

[2] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," vol. 42, no. 1, pp. 93–108, 2004..

[3] P . Clarkson and R. Rosenfeld, "Statistical language modeling using the cmu-cambridge toolkit," in Proceedings EUROSPEECH (N. F.G. Kokkinakis and E. Dermatas, eds.), vol. 1, (Rhodes, Greece), pp. 2707–2710, September 1997.

[4] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "A tree based statistical language model for natural language speech recognition," in Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, Issue 7, (Yorktown Heights, NY,USA), pp. 1001–1008, July 1989.