Pune Institute of Computer Technology Dhankawadi, Pune

A SEMINAR REPORT ON

NLG-Based Text Summarization

SUBMITTED BY

Name: Sania Edlabadkar

Roll No.: 314212 Class: TE-4

Under the Guidance of

Prof. Shital Girme

DEPARTMENT OF COMPUTER ENGINEERING Academic Year 2021-22

DEPARTMENT OF COMPUTER ENGINEERING Pune Institute of Computer Technology Dhankawadi, Pune-43

CERTIFICATE

This is to certify that the Seminar report entitled

"NLG-Based Text Summerization"

Submitted by

Sania Edlabadkar Roll No.: 31421

has satisfactorily completed a seminar report under the guidance of Prof. Shital Girme towards the partial fulfillment of third year Computer Engineering Semester I, Academic Year 2021-22 of Savitribai Phule Pune University.

Prof. Shital Girme Internal Guide

Dr. M. S. Takalikar
Head
Department of Computer Engineering

Place:Pune

Date: 09/11/2021

ACKNOWLEDGEMENT

It is my pleasure to present report on "NLG-Based Text Summarization". First of all, I would like to thank our Seminar Coordinator Prof. B.D.Zope, Head of Department Prof. M.S.Takalikar and Principal R.Shreemati for their encouragement and support.

I would also genuinely express my gratitude to my guide Prof. Manish Jansari, Department of Computer Engineering for his constant guidance and help. He has constantly supported me and has played crucial role in completion of this report. His motivation and encouragement from beginning till end to make this seminar a success.

Last but not the least I would thank all the faculty, my parents and friends who have helped me.

Contents

1	INT	INTRODUCTION		
2	MOTIVATIONS			2
3	A SURVEY ON PAPERS			3
	3.1	Text Summarization: A Review	.3	
	3.2 3.3	An Overview of Text Summarization Techniques	.3	
			.3	
	3.4	A Survey on NLP based Text Summarization for Summarizing Product Reviews	.3	
	3.5	Review on text summarization evaluation methods	.3	
4	FEA	TURES OF TEXT IN SUMMARIZATION		4
5	ME	THODOLOGIES		7
	5.1	The Abstractive Based Summarization Method	.7	
		5.1.1 Structured Based Approach	.7	
		5.1.2 Semantic Based Approach	.7	
	5.2	The Extractive Based Summarization Method	.9	
	5.3	Indicative Text Summarization	.10	
	5.4	Informative Text Summarization	.10	
6	INT	RINSIC AND EXTRINSIC EVALUATION TECHNIQUE	11	1
	6.1	Intrinsic Evaluation	.11	
		6.1.1 The utility method (UM)	.11	
		6.1.2 Content Similarity	.11	
		6.1.3 BLEU Scores	.11	
	6.2	Extrinsic Evaluation	.12	
		6.2.1 Shannon Game	.12	
		6.2.2 The Question Game	.12	
		6.2.3 Keyword Association	.12	
7	CO	NCLUSION	14	4
R	eferen	nces		15

NLG-Based Text Summarization

List of Tables

1	Features of text in summarization	5
2	Description Of Abstractive Summarization Approach	8
3	Advantages and Disadvantages of Abstractive Summarization Approach	9
4	Description Of Extractive Summarization Approach	11

Abstract

In this new era, where tremendous information is available on the internet, it is of most importance to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of large documents of text. Therefore, there is a problem of searching for relevant documents from the number of documents available and absorbing relevant information from it. In order to solve the above two problems, automatic text summarization is very much necessary.

Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings. There are two main approaches to the task of summarization—extraction, and abstraction. Extractive summarization is the strategy of concatenating extracts taken from a corpus into a summary, while abstractive summarization involves paraphrasing the corpus using novel sentences. The main aim of this review is to analyze the main approaches to automatic NLG(Natural Language Generation)-based text summarization based on their effectiveness and shortcomings and to review an application indetail.

This This paper also discusses various evaluation approaches on intrinsic and extrinsic techniques. In principle, text summarization is achieved because of the naturally occurring redundancy in text and because important (salient) information is spread irregularly in textual documents. Recognizing the redundancy is a challenge that hasn't been fully resolved yet.

Keywords

abstractive summarization, natural language processing, extractive summarization, natural language generation, intrinsic, extrinsic techniques

1 INTRODUCTION

There is an enormous amount of textual material, and it is only growing every single day. Think of the internet, comprised of web pages, news articles, status updates, blogs and so much more. The data is unstructured and the best that we can do to navigate it is to use search and skim the results. There is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for.

Text Summarization, is a topic that is related to the fields of philosophy and linguistics [2] are also included in the social sciences. The summary of large texts remains an open problem in natural language processing. Automatic Text Summarization is used to summarize large documents. Text summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document, to highlight the important parts of the text. The goal is to get the source, citing the most relevant information and showing the reader in a concise manner and in accordance with the needs of readers. Thus, this technology is expected to help the reader to absorb the information in the article summary/resume. Article shall be the length of their size, will lead the reader will be very difficult, if you want to read and understand all information in this article. Text synthesis will prepare a text which stores product/contain important parts of the original sentence.

Text summarizer can be classified based on input type: Single Document, where the input is small in textual context. Basic summarization models are built for such cases. Multi document, where the input can be comparatively long. The complexity increases here as more text leads to more semantic links being generated. Based on the aim, summarizer can be classified as Generic, where the model treats the input without any bias and prior knowledge. Domain-specific, where the model uses domain information to form a more accurate summary based on known facts. Query-based, where the summary only contains known answers to natural language questions about the input text. Based on output type summarizer can be classified as: Extractive, where important sentences are selected from the input text to form a summary. Techniques involve ranking the relevance of phrases in order to choose only those most relevant to the meaning of the source. Abstractive text summarization involves generating entirely new phrases and sentences to capture the meaning of the source document. This is a more challenging approach, but is also the approach ultimately used by humans. Classical methods operate by selecting and compressing content from the source document. Therefore, they are still far from reaching the human level, except for recent advances in the use of neural networks promoted by the advances of neural machine translation and sequencing models.

Applications of text summarizer are media monitoring, search marketing, internal document workflow, financial research, social media marketing, helping disabled people and more.

2 MOTIVATIONS

Propelled by modern technological innovations, data is to this century what oil was to the previous one. Today, our world is parachuted by the gathering and dissemination of huge amounts of data, which is mostly unstructured textual data. There is a requirement to develop automatic text summarization tools that allow people to get insights from them easily.

We enjoy quick access to enormous amounts of information. However, most of this information is redundant, insignificant, and may not convey the intended meaning. A lot of time is spent weeding out the unnecessary stuff before getting the information you want. Therefore, using automatic text summarizers capable of extracting useful information that leaves out inessential and insignificant data is becoming vital. There is a need to develop machine learning algorithms that can automatically shorten longer texts and deliver accurate summaries that can fluently pass the intended messages.

cm!

summarization can enhance the readability of documents, reduce the time spent in researching for information, and allow for more information to be fitted in a particular area.

3 A SURVEY ON PAPERS

3.1 Text Summarization: A Review

This paper takes into account the points of interest of both the extractive as well as abstractive methodologies longsi8de those systems utilized, their execution accomplished, alongside focal points and burdens of each methodology.

3.2 An Overview of Text Summarization Techniques

This survey paper covers extractive and abstractive summarization techniques. Summarization system should produce an effective summary in a short time with less redundancy having grammatically correct sentences. Both extractive and abstractive method yields good result according to the context in which they used. The reviewed literature opens up the challenging area for hybridization of these methods to produce informative, well compressed and readable summaries.

3.3 Automatic Extractive Text Summarization using K- Means Clustering

This work presents a method to achieve condensation of textual data, thus resolving the problem of redundancy and inaccuracy in a document. Proposed system performance is mediocre in comparison to commercial text summarizers. Through the course of evaluation, it is observed that k-means is a primitive clustering algorithm and advanced clustering algorithms are to be applied for superior results. Also, the process of sentence extraction from each 890 cluster must not be a random process but requires devising a specific algorithm.

3.4 A Survey on NLP based Text Summarization for Summarizing Product Reviews

Summarization has always been a necessity for many years as there is a huge amount of information being released on the internet every day. This paper described all the major summarizations techniques and the prominent works that are being done on each technique.

3.5 Review on text summarization evaluation methods

A general overview of automatic text summarization has been reviewed in this paper while focusing on various intrinsic and extrinsic evaluation techniques.

4 FEATURES OF TEXT IN SUMMARIZATION

Text Summarizers are utilized to distinguish the key sentences from the first content and afterwards those sentences are extracted and concatenated to form a summary. Lists of text summarization features are described below that are required to identify the key sentences from the original document.

Table 1: Features of text in summarization

Features	Description
Term-Frequency	Imperative terms given by various insights depend on the term recurrence. Notable terms are those terms that happen as often as possible or many numbers of times in a particular document. The score of the sentences increases as the frequency of any particular word increases. The widely used measure that is used to calculate the word frequency is TF IDF.
Title/Headline Word	In a summary, it becomes necessary to provide a suitable title or a heading for the particular summary that is produced from the original text or any document. So, the heading word should always be positively related to the summary. The words that are included as the title or the heading word should be related or should indicate the topic or the subject of the document.
Similarity	Similarity can be calculated byvarious expressive knowledge. Likeness between the sentence of the document as well as the heading of the document and the likeness between two sentences in a the document can be indicated by this text summarization feature.

Features	Description
Proximity	This feature is responsible for viewing the distance between the text units in the summary. It determines that sufficient distance should be there between two units so that their difference can be measured.
Sentence Length	This feature helps in maintaining the size of the summary. It should always be kept in view that excessive long or short sentences are not reasonable for the formation of a summary.
Cue Method	There are many positive and negative words that are present in the document from which the summary is to be extracted. By this method, a consequence of the constructive or the negative term used in the sentence is measured and the importance of the sentence is also measured. Many key words or cues are also there which can be used as "in summary", "in conclusion", "the paper describes", etc.
Thematic Word	The number of the thematic words in a sentence, this feature is important because terms that occur frequently in a document are probably related to the topic. The number of thematic words indicates the words with maximum possible relativity. We used the top 10 most frequent content words for consideration as thematic. The score for this feature is calculated as the ratio of the number of thematic words in a sentence over the sentence length.
Numeric Data	The number of numerical data in a sentence. A sentence that contains numerical data is important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of numerical data in a sentence over the sentence length.

NLG-Based Text Summarization

Features	Description
Proper Noun	In summarization of a document, the sentences having proper nouns are very important and should be considered as an essential statement to be included in the summary. example- the name of a person, place, or any other organization.
Location	Whether it is the first and last sentence in the paragraph, sentence position in the text gives the importance of the sentences. This feature can involve several items such as the position of a sentence in the document, section, paragraph, etc., proposed first and the last sentence highest ranking. The score for this feature: 1 for the first and last sentence, 0 for other sentences.

5 METHODOLOGIES

5.1 The Abstractive Based Summarization Method

Abstractive summarization generates a generalized summary by constructing new sentences alike a human being which is short and concise. Summary may contain new phrases that are not available in the source text. For generating abstractive summary language generation and compression techniques are necessary. Abstractive text summarization broadly classified into two types: Structure based and Semantic based approach.

5.1.1 Structured Based Approach

Structure based approach translates most important information from the document through cognitive schemas such as tree, ontology, lead and body phrase structure.

5.1.2 Semantic Based Approach

In Semantic based method, semantic representation of the document is given to natural language generation (NLG) system. This method focuses on identifying noun and verb phrases by processing linguistic data. These methods are Multi-modal semantic model, Information item based method, Semantic Graph Based Method.

Table 2: Description Of Abstractive Summarization Approach

Methods	Description
Tree Based Method	A dependency tree is used to represent the text of any to open documents. It uses some algorithm for the complete generation of any summary.
Template Based Method	Some kinds of templates are used in the representation of an entire document. Various patterns and the extraction rules are used in mapping to the template.
Ontology Based Method	This method uses the Ontology method also known as the knowledge-based a method so that the process of summarization can be improved. In many cases, Fuzzy Ontologies are also used in handling the uncertain data which cannot be handled by simple domain Ontology.
Lead and Body Phrase Method	In this method, various operations are performed on the phrases. Operations like insertion and substitution are performed. These are performed in order to rewrite the lead sentences in the summary.
Multi Modal Semantic Model	A model called a semantic model which keeps in view the concepts and the relationships among the concepts so the contents of the semantic model can be represented.
Information Item Based System	The content of the summary that is produced from this method is the generalization of the abstract representation of the original document. They are not generated from the sentences of the original documents.
Semantic Graph	Here, in this method the a summary is formed
Based Method	creating some rich semantic graphs.

Table 3: Advantages and Disadvantages Of Abstractive Summarization Approach

Methods	Advantages	Disadvantages
Tree Based Method	It usually works on some of the units of the document by which the summary can be read and understood easily.	Some of the modals lack in this method which is helpful in representing the summary abstractly.
Template Based Method	The kind of summary that is generated that is highly coherent because the summary that is produced is extracted using relevant items and information.	Designing of templates is required in this method and on the other hand, generalizing those templates are also very difficult.
Ontology Based Method	Due to this knowledge based model, drawing relations among the context becomes very easy.	In many cases, creating some of the rule-based methods can be very complex sometimes.
Lead and Body Phrase Method	Revision of the lead sentences can be done when this method is used in text summarization process.	Lacks some of the modals that is useful in representing the summary in the abstract form.
Multi Modal Semantic Model	This model produces an abstract summary because it has got an excellent coverage and the summary contains much graphical content.	The evaluation in this method is done manually.
Information Item Based System	It produces a concise summary that is richer in information and lesser in redundancy.	Due to the difficulty faced in creating meaning and grammatically correct sentences it is often rejected.
Semantic Graph Based Method	It produces summary that is concise and coherent and that are error free and redundant less.	It is limited to single document abstractive summerization.

5.2 The Extractive Based Summarization Method

The Extractive based summarization method selects informative sentences from the document as they exactly appear in source based on specific criteria to form summary. The main challenge before extractive summarization is to decide which sentences from the input document is significant and likely to be included in the summary. For this task, sentence scoring is employed based on features of sentences. It first, assigns a score to each sentence based on feature then rank sentences according to their score. Sentences with the highest score are likely to be included in final summary.

The extractive text summarization approach can be done in two steps:

Pre-Processing step — In this process, usually the original document is represented in the structured form. It mainly includes: Sentences Boundary Identification. Whenever a dot is present at the end of any particular sentence, then the sentence boundary is identified. Stop-Word Elimination — These are the common words in a text with no semantics. Stemming — The goal of stemming is obtaining the stemor the radix of each word, emphasizing its semantics.

Processing step — In this process, based on the relevance and the importance, the statements are selected from the original text, and then the selected sentences are added in the summary. Weights are also assigned to the sentences by using the weight learning method and then the final score of the sentences are calculated by using Feature-Weight equation and then the top ranked sentences are included in the summary.

Table 4: Description of Extractive Summarization Approach

Methods	Description
Term Frequency- (Inverse Document Frequency Method)	The recurrence of sentences is characterized by the quantity of the sentences in the record which contains that term. At that point vectors of the sentences are identified by the similitude of the question and also the most scoring sentences are picked to be a piece of the rundown.
Cluster-Based Method	It is obvious to imagine that synopses should always be addressed as distinctive "topics" appear in the reports. If the accumulation of the records for which outline is being given is of very surprising themes, archive bunching turns out to be relatively fundamental to create an important rundown. Sentence choice depends on the comparability of the sentences to the topic of the group (Ci). The following component is the area of the sentence in the report (Li). The last factor is its similitude to the main sentence in the report to which it has a place (Fi). Si=W1 * Ci + W2 * Fit W3 *Li Where, W1, W2, W3 are weightage for consideration in the rundown. The bunching k-implies calculation is connected.

NLG-Based Text Summarization

Text-Summarization using Neural Networks	This method includes preparation of the neural system to include in the sorts of sentences that should definitely be incorporated into our rundown. It uses a three-layered feed-forward neural system.
Graph Theoretic Approach	Chart theoretic portrayal of entries is given by a strategy for distinguishing the actual proof of topic after pre-handling steps, specifically evacuation of stop-words; sentences contained in the report aretaken as the main part in an undirected diagram.

5.3 Indicative Text Summarization

It doesn't contain informative content; it contains only metadata (a description of the document). It is used to inform the reader about the scope of the document and such to help help them decide whether or not to consult the original document. You can find such type of summaries, for example, on the verso of the title page of books and reports.

5.4 Informative Text Summarization

In this method, the main concept of the original document is first understood. After going through the complete document, the major purpose and the process should be understood and then the summary is to be formed completely in a natural language. It contains the informative part of the original text. After reading it, you can tell what are the main ideas in the original document. you can find informative summaries in research articles where the author tries to present the essential of its research.

6 INTRINSIC AND EXTRINSIC EVALUATION TECHNIQUE

A common way to evaluate the informativeness of automatic summaries is to compare them with human-made model summaries. First broad division for evaluating automatic text summarization systems is intrinsic and extrinsic evaluation methods [12]. An intrinsic evaluation tests the summarization system itself while an extrinsic evaluation tests the summarization based on how completion of some other task is affected. Intrinsic evaluations have assessed mainly the coherence and informativeness of summaries. Extrinsic evaluations, tested the impact of summarization on tasks like relevance assessment, reading comprehension, etc.

6.1 Intrinsic Evaluation

It measures the system in of itself, which is done by comparing to some old standard, (made by a reference summarization system or man-made using in- formants). Intrinsic evaluation mainly focuses on coherence and informativeness of summaries.

6.1.1 The utility method (UM)

The utility method (UM) allows reference summaries which consist of extraction units (sentences, paragraphs etc.) along with fuzzy membership in reference summary. The reference summary contains all the sentences of the source document(s) with confidence values for the inclusion in the summary. Furthermore, this method can be expanded to allow extraction units for exerting negative support on one another. This is predominantly useful when evaluating multi-document summaries.

6.1.2 Content Similarity

Here both extraction-based summaries and true abstracts can be applied to evaluate the semantic. One such is the Vocabulary Test (VT) where standard Information Retrieval methods are used to compare term frequency vectors calculated over stemmed or lemmatized summaries and reference summaries of some type.

6.1.3 BLEU Scores

Here an automatically computed accumulative n-gram matching scores (NAMS) between ideal summaries and system summaries is used as a performance indicator. Only content words were used in forming n-grams and n-gram matches between the summaries being compared.

6.2 Extrinsic Evaluation

Extrinsic evaluation on the other hand measures the efficiency and acceptability of the generated summaries in some tasks. If the summary contains some type of instructions, it is possible to measure at what extent it is possible to follow the instructions and the result. Proposed several games like scenarios at surface methods for summarization evaluation inspired by different disciplines. Among this include The Shannon Game (information theory), The Question Game (task performance), The Classification/ Categorization Game and Keyword Association (information retrieval).

6.2.1 Shannon Game

A variant of Shannon's measures in Information Theory is attempting to quantify information content by guessing the next token, e.g., letter or word, recreation of original text. The idea has been retrieved from Shannon's measures in Information Theory where three groups of informants to reconstruct important passages from the source article having seen either the full text, a generated summary, or no text at all. The information retention is measured in a number of keystrokes it takes to recreate the original passage. Hovy [24] has shown that there is a magnitude of difference across the three levels (about factor 10 between each group). The problem in Shannon's work is relative to the person doing the guessing and it is therefore implicitly conditioned on the reader's knowledge. Thus, information measure will infallibly change with more knowledge of the language, the domain, etc.

6.2.2 The Question Game

A variant of shanon's measures in Information Theory is attempting to qualify of the summary and an ability to convey key facts of the source article. This evaluation task is carried out in two steps. First the source articles are read by testers, marking central passages as they identify them. The testers then create questions corresponding to certain factual statements in the central passages. Next, assessors answer the questions 3 times: without seeing any document (baseline 1), after seeing a system generated summary, and after seeing the original document (baseline 2). A summary successfully conveys the key facts of the source article. It should be able to answer most questions, i.e. being closer to baseline 2 than baseline 1. This evaluation scheme has been used in the TIPSTER SUMMAC text summarization evaluation QA8 task, where [25] found an informativeness ratio of accuracy to compression of about 1.5.

6.2.3 Keyword Association

An inexpensive, shallower approach. It relies on keywords associated to the documents being summarized. For example, presented a human-judges with summaries generated by their summarization system together with five lists of keywords taken from the source article as presented in the publication journal. The judges were given the task to associate each summary with the correct list of keywords. If successful, the summary is said to cover the central aspects of the article since the keywords associated to the article by the publisher were content indicative. Major advantage is that it requires no cumbersome manual annotation.

7 CONCLUSION

The status, and state, of automatic summarizing, have radically changed through the years. The biggest challenge for text summarization is to summarize content from a number of textual and semi-structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user. Text summarization software should produce an effective summary in less time and with the least redundancy. This paper takes into account the points of interest of both the extractive as well as abstractive methodologies alongside those systems utilized, their execution accomplished, alongside focal points and burdens of each methodology. As abstractive synopsis requires additional learning and thinking, it is a bit complex than extractive methodology at the same time, abstractive rundown gives a more important and fitting outline contrast with the extractive approach. Summaries can be evaluated using intrinsic or extrinsic measures. While intrinsic methods attempt to measure summary quality using human evaluation, extrinsic methods measure the same through a task-based performance measure such as an information retrieval-oriented task. Research in this field will continue due to the fact that the text summarization task has not been finished yet and there is still much effort to do, investigate and improve.

References

- [1]K. Shetty and J. S. Kallimani, "Automatic extractive text summarization using K-means clustering," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEEC-COT), 2017, pp. 1-9, doi: 10.1109/ICEECCOT.2017.8284627.
- [2] Ganesh, Prakhar and Saket Dingliwal. "Restructuring Conversations using Discourse Relations for Zero-shot Abstractive Dialogue Summarization." arXiv: Computation and Language (2019): n. pag.
- [3]Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez and Krys Kochut, "Text Summarization Techniques: A Brief Survey" International Journal of Advanced Computer Science and Applications(IJACSA), 8(10), 2017.
- [4]S. Biswas, R. Rautray, R. Dash, "Text Summarization: A Review," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), 2018, pp. 231-235, doi: 10.1109/ICDSBA.2018.00048.
- [5]R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.
- [6] Andhale, Narendra Bewoor, Laxmi. (2016). An overview of Text Summarization techniques. 1-7. 10.1109/ICCUBEA.2016.7860024.
- [7] Prathima, M H R, Divakar. (2018). Automatic Extractive Text Summarization Using K-Means Clustering. International Journal of Computer Sciences and Engineering. 6. 782-787. 10.26438/ijcse/v6i6.782787.
- [8] Xu, Shusheng, Xingxing Zhang, Yi Wu, Furu Wei and Ming Zhou. "Unsuper-vised Extractive Summarization by Pre-training Hierarchical Transformers." FINDINGS (2020).
- [9]Indu, M; Kavitha, K V (2016). [IEEE 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS) Bangalore, India (2016.5.6-2016.5.7)] 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS) Review on text summarization evaluation methods. , (), 1–4. doi:10.1109/RAINS.2016.7764406

P:F-SMR-UG/08/R0