Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 6, Issue 4, April-2019

Abstractive Text Summarization of Research Articles using RNN

Professor Dr. S.Saraswathi
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
swathi@pec.edu

S.Savitha
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
savidharman98@gmail.com

S.Gayathri
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
gayathripec98@gmail.com

T.P.Rahul
Department of Information Technology
Pondicherry Engineering College
Puducherry, India
rahultp37@gmail.com

Abstract*— *Abstractive summarization methods aims for producing important information using new way. In other words, they use advanced natural language techniques to interpret and examine the text in order to generate a new shorter text that conveys the most marginal information from the input text. This information overload increases in great demand for more capable and dynamic text summarizers. There are a variety of applications like summaries of articles from newspaper, textbook, educational magazine, anecdotes on the same topic, event, research paper, weather report, stock exchange, CV, music, plays, film and speech. Hence it finds its importance. In this paper we discuss the use abstractive summarization for research papers using RNN LSTM algorithm.

Keywords*—*Abstractive text summarization; RNN; sequence to sequence; text mining; GloVe;

1. INTRODUCTION

Text summarization is a process of providing crisp and important information from input documents and presents that information in the form of summary. In recent years, need for summarization can be seen in various purpose and in many domains such as news articles summary, email summary, short message of news on mobile, research articles and information summary for businessman, government officials, researchers online search through search engine to receive the summary of relevant pages found, medical field for tracking patient's medical history for further treatment.

The amount of printed data increases the need for automatic summarization. A huge amount of data is accessible on the web, however, to deal with the required data is an endless task. The aim of document summarization is to extract the consolidated adaptation of the initial document [1]. An overview of the document is valuable because it can provide an analysis of the initial document in less time. Readers can choose whether or not to pursue the full archive in the wake of spending the outline. For instance, before reading the full paper initially readers will read the abstract of a logical article. Web indexes additionally utilizing synopses of content to help clients settle on significant choices.

Automatic text summarization is the task of producing a concise and fluent summary which is used for preserving critical content and overall meaningful information. Numerous approaches have been introduced over the last decade for automatic text summarization and they have been applied in variety of areas of domains. In the scenario of search engines, a jest of the document is represented by snippets. Other examples include news websites which produce brief descriptions of news topics usually as headlines to make browsing or knowledge extractive approaches easier. Automatic text summarization is very challenging, because when we manually summarize a piece of text, we usually consume a lot of time read it entirely to comprehend, and then

All Rights Reserved, @IJAREST-2019

generate a summary containing the key points. Since computers lack human intelligence and vocabulary, it makes automatic text summarization a very difficult and non-trivial task.

The aim of automatic text summarization is to make large text into short and readable format. The automatic summarization of text is a popular approach in the domain of natural language processing (NLP). Using sentence extraction and statistical analysis, significant achievements in text summarization have been obtained [2]. The task of Multi-Document Summarization(MDS) is to produce concise and crisp summary which gives important information about the given set of document. Multi Document Summarization would be helpful for the users to quickly comprehend the primary idea of document collection, and it has been shown that multi document summarization could also be used to improve the performance of information retrieval systems. MDSs can be categorized in two classes. First one is query-based Multi Document Summarization where the generated summary is biased according to the given query. The category of topic-focused MDS is where the summary is mainly focused on a specific topic or category.

Abstractive summarization systems rephrase or uses words that were not in the original text to generate new phrases. Abstractive approaches are naturally harder [3]. When the model understands the document and tries to generate a short description using the understanding then it is called a perfect abstractive summary. Much harder than extractive. Has complex capabilities like generalization, paraphrasing and incorporating real-world knowledge.

2. LITERATURE SURVEY

2.1 Tree based summarization method

In tree based summarization we use a dependency tree. This tree is used to represent the contents of a document. A language generator or an algorithm is used for generation of summary in this method. The approach proposed in [4] automatically fuse similar sentences across news articles on the same event. Language generation gives concise summary hence it is used in this method. In this approach, first the dependency trees are obtained by analysing the sentences. A basis tree is set by finding the centroid of the dependency trees. It next augments the basis tree with the sub-trees in other sentences and finally prunes the predefined constituents.

Limitation: It lacks a complete model which would include an abstract representation for content selection.

.

2.2 Template based summarization method

This technique represents a whole document in the form of templates. Linguistic patterns or extraction rules are queried to find text snippets which are indicators of the summary content that will be mapped into template slots using this. Using CICERO Information extraction(IE) system the proposed approach produces summaries of many newspaper documents which presents a multi document summarization system. To extract information from multiple documents, CICERO requires a template representation of the topic from multiple documents. The templates are filled with mandatory text snippets which were extracted from Information Extraction systems. An illuminating advantage of this approach is that the generated summary is highly coherent because it relies on information identified by IE system.

Limitation: It cannot handle the task if multi document summarization requires information about similarities and differences between a number of documents.

2.3 Ontology based summarization method

Majority of the documents on the web are domain related because. Every domain has a unique knowledge structure. They can be better represented by ontology. The fuzzy ontology with the use of fuzzy concepts was introduced for Chinese news summarization [6] to provide a better description of domain. The domain ontology for news events will be defined by the domain experts. Followed by, the document pre-processing phase produces the meaningful terms from the news corpus and the Chinese news dictionary. Then, the events of news are classified into meaningful terms using term classifier. Each of the fuzzy concept of the fuzzy ontology, the fuzzy inference phase produces the membership degrees which is associated with various events of the domain ontology. The advantage of this approach is that it exploits fuzzy ontology to deal with data that is uncertain.

Limitation: Ontology based summarization approach is limited to Chinese news, and might not be applicable to English news

All Rights Reserved, @IJAREST-2019

2.4 Lead and Body Phrase based summarization

This approach is based on the functions of phrases inclusion and replacement for the purpose of rewriting the lead sentence. Allied literature of this approach is conferred as follows. An abstractive approach proposed by [7] revise lead sentences in a news broadcast. Co-reference relation of noun phrases (NPs)is not made use by this approach. In this approach, first the same triggers or chunks are explored in the lead and body sentences. Then, maximum phrases (revision candidates) of every chunks are recognised and associated using similarity metric. Switch of phrase of the body for the lead phrase takes place if body phrase has required phrase in the lead and body phrase is richer in information. Inclusion of body phrase into the lead sentence, if a body phrase has no equivalent in the lead sentence. The advantage of this method is that it found semantically appropriate revisions for revising a lead sentence.

Limitation: Parsing errors degrade sentential completeness such as grammaticality and repetition. Second of all, lead and body phrase summarization concentrates on rewriting techniques, and surpluses a model which would be inclusive of an abstract representation for selection of content.

2.5 Rule Based Summarization method

This approach in [8] produces crisp abstractive summaries from clusters of news articles on same event. The abstraction approach uses a rule based information extraction module, content selection heuristics and several patterns for sentence development. In order to generate extraction rules for abstraction scheme, many verbs and nouns having similar meaning are determined. The syntactic position of roles is also identified. The information extraction (IE) module identifies many candidate rules for each aspect of the category. Summary is generated using generation patterns. These patterns are designed for each abstraction scheme. The advantage of this method is that it has a potential for creating summaries with greater information density.

Limitation: The main drawback of rule based methodology is that all the rules and patterns are manually written, which is tiresome and consumes a lot of time.

2.6 Multimodal Semantic Model

In this method, a semantic model, is built to represent the contents (text and images) of multimodal documents which is used to abduct concepts and relationship among concepts. The concepts which are important are rated based on some measure and lastly the concepts chosen are expressed as sentences to form summary. In [9], a framework was proposed for generating an abstractive summary from a semantic model of a multimodal document. Multimodal document contains both text and images. The framework has three steps: In first step, a semantic model is constructed using knowledge representation based on objects (concepts) standardized by ontology. In second step, information density metric is used to rate the information content. The metric regulates the applicability of concepts based on fullness of attributes, the relationship with other concepts is numbered and the number of expressions which show the frequency of concept in the current document. In third step, the important concepts are expressed as sentences. A semantic model stores the expressions that are observed by the parser for expressing concepts and relationship. A major highlight of this framework is that it generates abstract summary, it has a coverage which is excellent since it is inclusive of pertinent textual and graphical content from the whole document.

Limitation The limitation of multimodal framework is that it is manually evaluated by humans. An automatic evaluation of the framework is desirable.

2.7 Information Item Based Model

In this method, the contents of summary are generated from information item of source. In sentence generation module, a sentence is directly generated from INIT using a language generator, the NLG realizer Simple NLG [10]. Sentence selection module ranks the sentences based on their average Document Frequency (DF) score. Finally, a summary generation step produces the highly ranked generated sentences with dates and locations.

Limitation: Many candidate information items are rejected due to the difficulty of creating meaningful and grammatical sentences from them. Secondly, linguistic quality of summaries is very low due to incorrect parses.

All Rights Reserved, @IJAREST-2019

2.8 Semantic Graph Based Method

It summaries creating Rich Semantic Graph (RSG) which is a semantic graph. The final abstractive summary is achieved by reducing the generated semantic graph. The abstractive approach proposed by [11] consists of three phases as shown in figure 1. In the first phase rich semantic graph is used to represent input document. In the second phase the heuristic rules are applied to reduce the rich semantic graph. Finally, the third Phase generates the abstractive summary from the reduced rich semantic graph and generates the summarized text.

Limitations: Semantic graph based method is limited to single document abstractive summarization.

3. OVERVIEW OF SYSTEM

3.1 Problem Definition

Understanding the relation between various research article is one of the main difficulties faced by the researchers. To make the process easier we make use of the method of abstractive text summarization for better understanding. Abstractive text summarization derives a summary of the given research articles. The summary will give the title, author names, year of publication, technique used to solve the problem defined, dataset used and the parameters used for evaluation.

3.2 System Model

At first the given input research articles are converted to JSON format. From the Json file of each research article the metadata of the research articles are extracted. Then we parse through the article and extract sentences related to the metadata of the research article. Then we perform the pre-processing of the given input research articles, that is the removal of the stop words and the stemming words in the research articles. A word that is commonly used such as “the”, “a”, “an”, “in” in a search engine are called as the stop words, these stop words has been programmed to ignore, both when indexing entries for processing and when retrieving them as the result of a search query. Afterwards, we utilize POS tagger [40] to assign Part of Speech (POS) tags to different tokens of the semantic arguments. The POS tags consists of verb(V), noun(NN), adverb(RB)) and adjective (JJ) etc. From the tagged words extract the keywords. Once the keywords are extracted the sentences containing the keywords are selected and a sentence score is assigned. GloVe is used to create word embedding. The encoder captures the meaningful source of document. Hence the encoder is where the complexity of the system resides. The encoder maps a given word to a vector representation. The decoder must convert the word embedding that is the vector representation to a word that gives the same meaning as the vectorised word. The decoder must generate each word in the output sequence given by context vector and generated sequence.

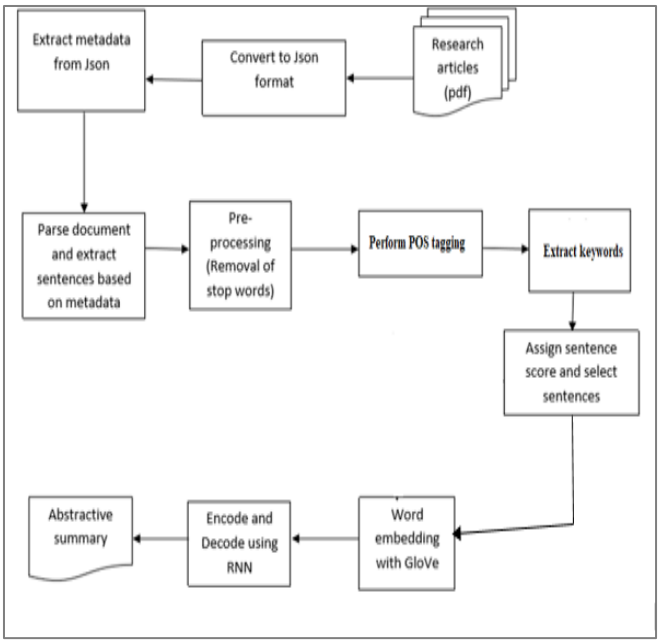


Figure. 3.2.1 Architecture Diagram for Abstractive text summarization

All Rights Reserved, @IJAREST-2019

3.3 Encoder Decoder RNN LSTM

This architecture is has two models. One model is used for reading the given input sequence and encodes it into a fixed length vector. The second model is used for decoding the fixed length that is output of the encoder and gives a predicted sequence as the output. There are two recurrent neural networks in RNN Encoder Decoder that acts as encoder decoder. The main task of the encoder is to map the variable length source into a fixed length vector. The decoder maps the vector representation back to the variable length target sequence.The LSTM was mainly designed for natural language processing problems. The RNN encoder and decoder takes into account both semantic and syntactic structures of the phrases. The encoder is given the input sequence one word at a time. The words are then passed through the embedding layer where the words get transformed into a distributed representation. A multi-layer neural network is used to represent the distributed representation. The encoder is single layer bidirectional LSTM. The decoder generates each word in the output sequence as the context vector and generated sequence. The context vector is the encoded representation of the source document. The generated sequence is the word or the sequence of words that is generated as summary.

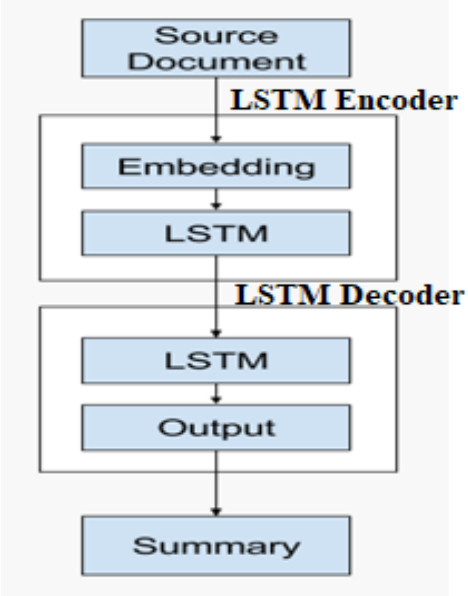


Figure. 3.3.1. Architecture of LSTM RNN

3.4 Input & Output

Input- Research articles of the domain machine learning.
Output – Summary of the given research articles that gives the title, author name, year of publication, technique used in the paper for implementation, dataset used and the parameters used for evaluation.

3.5 Modules Description

3.5.1 Conversion to Json format and extraction of metadata

The pdf file that is the research article that is given as input is converted to Json format. From the Json file the metadata information such as title, author and the keywords used in the article can be extracted.

3.5.2 Pre Processing Module

Using the keywords extracted from the Json file the whole pdf is parsed and sentences based on the metadata are extracted. This module is used for pre-processing the sentences extracted from input documents based on the metadata i.e. to remove the stop words and stem words in the document.

All Rights Reserved, @IJAREST-2019

3.5.3 Tagging and Sentence Selection Module

After the pre-processing, the sentences are tagged using POS tagger. We utilize POS tagger to assign Part of Speech (POS) tags to different tokens of the semantic arguments. The POS tags consists of verb(V), noun(NN), adverb(RB)) and adjective (JJ) etc. From the tagged words extract the keywords. Once the keywords are extracted the sentences containing the keywords are selected and a sentence score is assigned.

3.5.4 Summary Generation Module

GloVe is used to create word embedding. The encoder is where the complexity of the model resides as it is responsible for capturing the meaning of the source document. The decoder must generate each word in the output sequence given by context vector and generated sequence. The sequence gives the generated abstractive summary.

3.6 Result analysis

Document summarization can be measure using two parameters namely Precision and Recall. The several aspects which are used for of text(linguistic) quality:
Grammaticality: Non-textual items must not be used in the text (i.e., markers) or production errors or incorrect words.
Non-redundancy: There should not be repeated content in the document.
Reference clarity: In the summary, we must refer nouns and pronouns clearly.
Coherence and Structure: There should be good structure for the summary and the sentences must be coherent

3.6.1 Mean Opinion score

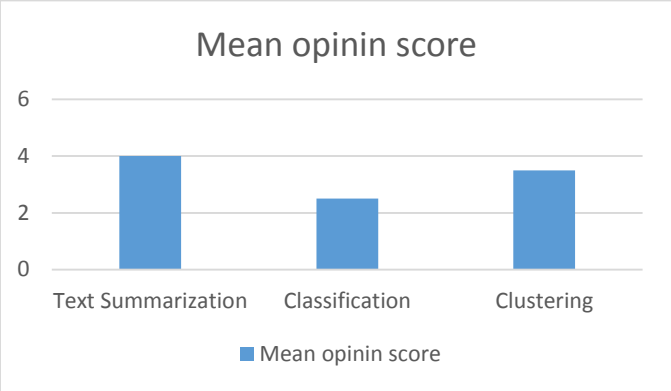
A Mean Opinion Score (MOS*)* is a numerical measure which is given by human experts who use the system and give a score between 0-5. The value of each score is explained below:

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Sightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Mean opinion score was collected from 100 people for each domain and their average was calculated.

DOMAIN	NO OF PEOPLE WHO GAVE SCORE AS					AVERAGE SCORE
	5	4	3	2	1	
TEXT SUMMARIZATION	30	40	20	7	3	4
TEXT CLASSIFICATION	7	20	60	8	5	3
DOCUMENT CLUSTERING	10	40	35	8	7	3.5

All Rights Reserved, @IJAREST-2019



4. CONCLUSION

In our proposed work, we explore the problem of abstractive text summarization of research articles using RNN LSTM algorithm which resolves the difficulty of the readability of the research articles.

REFERENCES

[1] Atif Khan,Naomie Salim,Haleem Farman,Murad Khan,Bilal Jan,Awais Ahmad,Imra, Anand Paul, "Abstractive Text Summarization based on Improved Semantic Graph Approach", International journal of parallel programming,02 February 2018.

[2] Y. Song, S. Pan, S. Liu, F. Wei, M. X. Zhou and W. Qian, "Constrained Text Coclustering with Supervised and Unsupervised Constraints," IEEE Transactions Journal on Knowledge and Data Engineering, (Volume:25, Issue: 6), pp. 1227-1239, 2013

[3] Y. Wang, X. Ni, J.-T. Sun, Y. Tong and Z. Chen, "Representing Document as Dependency Graph for document clustering" in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011.

[4] S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with gistexter," in Document Understanding Conferences, 2002.

[5] C.-S. Lee, et al., "A fuzzy ontology and its application to news summarization," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol.35, pp. 859-880, 2005.

[6] H. Tanaka, et al., "Syntax-driven sentence revision for broadcast news summarization," in Proceedings of the 2009 Workshop on Language Generation and Summarisation, 2009, pp. 39-47.

[7] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 2012, pp. 354-358.

[8] H. Saggion and G. Lapalme, "Generating indicative-informative summaries with sumUM," Computational Linguistics, vol.28, pp. 497-526, 2002.

[9] C. F. Greenbacker, "Towards a framework for abstractive summarization of multimodal documents," ACL HLT 2011, p. 75, 2011.

[10] A. Gatt and E. Reiter, "SimpleNLG: A realisation engine for practical applications," in Proceedings of the 12th European Workshop on Natural Language Generation, 2009, pp. 90-93.

[11] I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization," in Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on,2012, pp. 132-138.

All Rights Reserved, @IJAREST-2019