Grenze International Journal of Engineering and Technology, Special Issue

Abstractive Summarization using AMR and Random Walk Algorithm

Payal Sancheti [1], Rajashree Shedge [2] and Namita Pulgam[3]
[1-3]Ramrao Adik Institute of Technology / Dept. of Computer Engineering, Navi Mumbai, India
Email: jainpayal2494@gmail.com, rajashree.shedge@gmail.com
ashwinigudewar@gmail.com

***Abstract*— The information on Internet is increasing massively in amount. From such enormous availability of data, fetching meaningful knowledge in short span of time is too taxing. With this information deluge, need of automatic summarization has increased twofold. Text summarization is an approach of condensing longer texts without altering it's meaning and is classified into Extractive and Abstractive Summarization. Former technique reformulates text by extracting related information and concatenating into final summary similar to copy-paste operations. More emphasis is given on sentence structure that is not sufficient to capture gist of information required. The latter makes use of novel words and connectors in forming summary from the aggregated information. The methods of abstractive summarization are less pertinent and tend to ignore semantics. One of the existing techniques, which have graphical approach towards summarization, is Abstract meaning representation (AMR), which is Standard Knowledge framework but ignores identical sentences. The proposed method will build a rooted labeled graph of sentences, checks the similarity of concepts and improves the existing system resulting in fluent summaries.**

***Index Terms*— Text Summarization, Abstractive summarization, Abstract Meaning Representation, random walk algorithm.**

I. INTRODUCTION

Natural Language Processing (NLP) is a field of study that focuses on the interaction between humans and machines. The challenges associated with NLP are retrieval and processing of large amount of information, which is in most cases, real-time. One such important process in Information Extraction (IE) is Text Summarization. It basically comprises of condensing abundant information into a precise form by discarding generating algorithms [1]. Furthermore, the actual time required to read the text is greatly reduced when reading summaries. Current research focuses more on extracting information, i.e. summarizing information based on sentence-extraction. This process is much simpler as compared to creating abstract. Moreover, more emphasis is given to sentence structure here to enable the summary document to be read as is. Hence, creating words and sentences, which helps to capture gist, is not sufficient. The scope of this paper limits itself to finding new approaches that can create summaries based on the approaches used by humans when creating the same. There are two approaches of summarization mainly: Extractive and Abstractive.
The extractive summarization focuses more on extracting the exact information the text wishes to convey [2],

Grenze ID: 01.GIJET.4.3.1

whereas abstractive approach focuses more on expressing what the text wants to convey. In other words, sentences and additional information may be created to convey the information. The organization of paper as follows: Related work de-scribed in section II. Section III defines the problem. Proposal is discussed in section IV and it is finally concluded in section V.

II. Literature Survey

The authors Mehdi Allahyari, Seyedamin Pouriyeh and Mehdi Assefi focussed on various extractive approaches for single and multi-document summarization. Also highlighted recent trends and progress in text summarization like techniques for representing topics, methods involving frequency of words, graph-based and machine learning techniques [3]. The first approach of summarization is extractive [4] that was used for automatic summarization involved calculating frequencies of terms in a sentence and rating sentences. Sentences which ranked higher i.e. had more frequently appearing terms were considered important and included in summary. Extractive summarization has been given utmost importance in many of the research studies, the approach being selecting sentences from documents [5]. To find keywords and phrases, mostly statistical methods are used [6].For larger and complex documents, which require lot of sentence restructuring and including information from full text, the extractive summarization methods cannot create summaries as accurate as created by humans [7]. Abstract summaries can be created by two ways in general, Syntactic (or Linguistic) and Semantic approach. Syntactic approaches use a syntax parser, which can be used to understand the context of the document to create summary [8-9]. Regina Barzilay and Kathleen R. McKeown proposed concept of sentence fusion, which basically involves aligning syntactic tress of input. This not only enables to summarize information but also helps in including information common in most sentences. However, this method had no selection criteria for sentences in summary. This led to summaries not being as coherent as the input text[10]. A template representation of a topic can be viewed as a list of semantic roles, each role being a slot that is filled by information extracted from text. In [11-12] refined a corpus-based technique which created an extempore semantic roles of collected topics for representing template also enhanced multi-document summarization by addition of statements from multiple documents based on their mapping from the template slots provided good results for creating variable size summaries. In [13], Information extraction rules, content selection and generation of patterns such abstraction scheme were made not by machine, but by hand. All these rules were designed to address a subcategory or a theme. Thus, these rules that clips information from the corpus helps to produce similar data, which created summaries. From sentence fusion method authors in [14] obtained an idea of adding phrases of syntactic type for which they made use of insertion and substitution from the phrases modifying the identical head chunks in forefront sentences. An innovative approach proposed by Xiang Zhang [15] for automatic ontology summarization was based on RDF Sentence Graph. The authors compared different centrality measurements in evaluating the salience of RDF sentence and defined a reward-penalty re-ranking algorithm to make the understandable summaries. The performance of all centralities produced qualified summaries after re-ranking and shown promising results. The authors in [16-17] aims to advance the state of the-art in automatic summarization by generating abstractive summaries from a semantic model built from the original article. Important graphical matter which are potentially ignored as they are non text components, so by including concepts from information graphics, produced unified summaries resulting covering the abstract idea of the entire document. In [18] a novel approach using rich semantic graph (RSG) reducing technique helped to create abstractive summary for a single document. Firstly, this method graphically represents the document, then reducing the creating the RSG and finally generating summary from the condensed graph. Banarescu et al [19] as first introduced AMR. It focuses on capturing the meaning of the text, by giving a specific meaning representation to the text about who is doing what to whom in a sentence. This graph theory symbolizes to render similar appearance of sentences, which have the equivalent fundamental meaning. Research work on summarization using AMR was initiated by Liu et al [20]. He introduced a statistical summarization model directed by Abstract Meaning Representation. The approach suggested a structured prediction algorithm that converts semantic graphs of the input into single summary semantic graphs. At that time, due to unavailability of text generator from AMR, their work was limited till extracting the summary graph. This method extracts a single summary graph from the story graph. Extractive Summarization is easy to build as it extracts sentences from the corpus and adds it to summary; therefore it fails to convey the gist of overall article. In summary, using grammatical structure instead of sentence analysis fairly limits our capabilities to find out more meaningful relationships in text. This literature survey focuses on how extractive methods fail to convey the actual meaning of the text without actually forming meaningful context. On the other hand, abstract generation methods focus more on the semantics and give us the refined summary. The biggest task in abstract generation is keeping the semantics of the corpus intact while ensuring the entire context is covered in the summary. This problem is outlined in problem statement with an approach made to reach the solution.

III. PROBLEM STATEMENT

Structured and Semantic based Abstractive approaches mainly focus to produce coherent summary by understanding meaning and ideas discussed in paper. It also performs deeper analysis to form complete, compact summaries. The extractive approach fails to consider the semantic representation of text; whereas the other examines the semantic representation of documents like multimodal and information item based techniques are time consuming and fail to provide proper relationship between sentences. Therefore there is need to create a system that properly analyses the input documents and create summary that are identical to summary spawned by humans. The research work will try to accomplish the above-mentioned need and for that graph theory will be applied for representing sentence structure as well as relationship between them. AMR is a combination of sentence-semantic representation. These graphs have concepts as its nodes and edges are its relationships, which fail to consider similar sentences. Therefore there is need to consider similarity between concepts which helps to produce distinct graphs and generate customized summary.

IV. PROPOSAL

A. Proposed Framework

An appropriate framework in order to represent textual information, which is capable to allow machines to execute language, perform analysis and generate summary will be pivotal to achieve the task of summarization. "Fig. 1" describes the flow of proposed system:

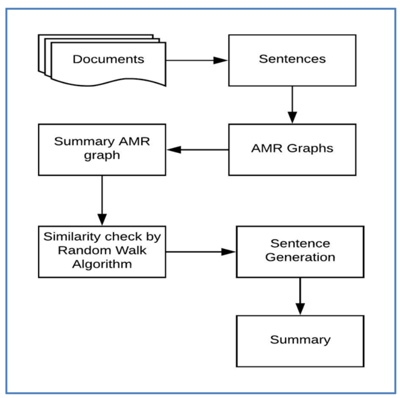

Figure 1. Proposed Framework

For the input dataset, CNN-Daily mail corpus will be used for processing. It is well suited for this task as the average summary size is around 3 or 4 sentences. This dataset has around 300k document summary pairs with stories having 39 sentences on average [22].

1) <u>Sentences:</u> This step separates the document into sentences. It makes use of end-of-sentence punctuation marks, question marks, exclamation and periods.

2) <u>AMR graphs:</u> The next step is conversion of story sentences to Abstract Meaning Representation. JAMR-Parser version 2 is used as its openly available. Flanigan et al [21] proposed a Discriminative graph based parser called JAMR. It makes use of annotated AMR corpus to train its model parameters. The process mainly has two main subtasks: Identifying Concept nodes using a semi Markov model that is solved using Viterbi algorithm and Identifying relation between nodes by searching a maximum spanning connected sub graph from among all possible relations that can exist between concepts.

3) <u>Summary AMR graphs:</u> After parsing the output will be AMR graphs for the story sentences. Now in this step, the system will extract the AMR graphs of the summary sentences using story sentence AMRs. The division of task in two parts: First is finding the important sentences from the story and then extracting the key information from those sentences using their AMR graphs. Important sentences are found by first-n method, where n stands for number of sentences. [22]

4) Similarity check by Random Walk Algorithm: The graph will be mapped into matrix to form a*a probability matrix Q, a forms the no. of nodes in graph. Each element of matrix Q will be normalized. Forming an initialization matrix for random walk such as $A_0 (X_i) = 1$, if walk starts from i^{th} node, else 0.
Algorithm:

(i) Let $A = A_0$

(ii) A (new) $= \alpha * Q * A + (1 - \alpha) * A_0$; ($\alpha$ = 0:5)

(iii) A = A (new)

(iv) Repeat step 2 and 3 until A (new) reaches stable position or beyond threshold

Through multiple iteration, a stable probability distribution matrix will express association strength between initial node and other nodes.

5) Sentence generation: From the extracted AMR graphs, sentences will be generated with the help of available Neural AMR generator [22].

B. Performance Evaluation

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) consists of measures to evaluate the predicted summary by comparing it to other summaries generated by humans. It assess the summaries by: counting the number of overlapping units such as word pairs, n-grams and sequences of words the predicted summary to be evaluated against target summaries created by human. The metrics to evaluate the system will use precision and recall using the overlap to get quantitative value.

Recall = No. Of overlapping words / total No. Of words in reference summary (1)

Precision = No. Of overlapping words / total No. Of words in system summary (2)

It is expected that with the help of ROUGE metrics summary generated from existing system will be fairly evaluated against summary generated by proposed system.

V. CONCLUSION

The paper provides a crux of extractive and abstractive summarization. The study shows that methods of former approaches fail to consider the semantics of the sentences which is taken care in graphical approach of latter technique. AMR helps to build proper internal semantic representation but ignores similarity of statements, which will be useful in forming coherent summaries. The proposed method enforces to check similar concepts and improves the final summary by contributing another set of weights to the existing system of weights for the selected sentences. Rogue measures will be helpful in providing quantitative value, which will help to compare the predicted and target summaries.

REFERENCES

[1] Kallimani, J. S., K.G.Srinivasa, R. B. Eswara, “Information Extraction by an Abstractive Text Summarization for an Indian Regional Language ”. Proceedings of IEEE NLP-KE, Tokushima, Japan, pp. 27-29, November 2011.

[2] Pierre-Etienne Genest, Guy Lapalme and RALI-DIRO, “Framework for Abstractive Summarization using Text-to-Text Generation”. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 64-73, 24, June 2011.

[3] Mehdi Allahyari, Seyedamin Pouriyeh and Mehdi Assefi, “Text Summarization Techniques: A Brief Survey”, arXiv, USA, July 2017.

[4] Luhn and H.P, “The automatic creation of literature abstracts”, In IBM J. Res. Dev. pp. 159-165, 1958.

[5] Kupiec, J.Pedersen and J.Chen, “A trainable document summarizer”. In Proceedings of the 18th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, Washington, pp. 68-73, 1995.

[6] Knight and K.Marcu, “Statistics-based summarization-step one: Sentence compression”, In Proceedings of the National Conference on Artificial Intelligence, AAAI Press, Menlo Park, pp. 703-710, 2000.

[7] Kanitha D.K and D. Muhammad Noorul Mubarak, “An Overview of Extractive based Automatic Text Summarization Systems”, In International Journal of Computer Science Information Technology (IJCSIT), vol. 8, no. 5, October 2016.

[8] Tanaka, H., Kinoshita, Kobayakawa, and Kato, N., “Syntax-driven sentence revision for broadcast news summarization”, In Proceedings of the 2009 Workshop on Language Generation and Summarisation, Singapore, pp. 39-47, 2009.

[9] Genest and Lapalme, “Framework for abstractive summarization using text-to-text generation”, In Proceedings of the workshop on monolingual text-to-text generation, Oregon, USA, pp. 64-73, 2011.

[10] Regina Barzilay and Kathleen R. McKeown, "Sentence Fusion for Multi Document News Summarization", In Computer Science and Artificial Intelligence Laboratory Computational Linguistics, vol. 31, no. 3, 2005.
[11] Sanda M. Harabagiu and Finley Lacatusu, "Generating Single and Multi-Document Summaries with GISTEXTER", In U. Hahn D.Harman (Eds.), Proceedings of the workshop on automatic summarization, 2005.
[12] Regina Barzilay, Kathleen R. McKeown and Michael Elhadad, "Information Fusion in the Context of Multi-Document Summarization", In Proceedings of the 37th annual meeting of the Association for Computational Linguistics, 1999.
[13] Genest, Pierre-Etienne, and Guy Lapalme, "Fully abstractive approach to guided summarization." In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, vol. 2, 2012.
[14] Hideki Tanaka, Tadashi Kumano and Kato, "Syntax-Driven Sentence Revision for Broadcast News Summarization", In Proceedings of the 2009 Workshop on Language Generation and Summarization, Singapore, pp. 39-47, 6 August 2009.
[15] Zhang, X. and Cheng, "Ontology Summarization Based on RDF Sentence Graph", In 16th Inter. World Wide Web Conference Banff, Alberta, Canada, May 8-12, 2007.
[16] UzZaman, Naushad, Bigham, and Allen, "Multi-modal summarization of complex sentences.", In Proceedings of the 16th International conference on Intelligent user interfaces, ACM, 2011.
[17] Greenbacker and Charles F, "Towards a framework for abstractive summarization of multimodal documents." In Proceedings of the ACL Student Session, 2011.
[18] Moawad, Ibrahim F., and Mostafa Aref, "Semantic graph reduction approach for abstractive Text Summarization." In Computer Engineering Systems (ICCES),Seventh International Conference IEEE, 2012.
[19] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Martha Palmer, Philipp Koehn, and Nathan Schneider, "Abstract meaning representation for sembanking", In Proceedings of the 7th Linguistic Annotation Workshop Interoperability with Discourse, pp. 178-186, Aug 8-9, 2013.
[20] Fei Liu, Flanigan Jeffrey, Thomson Sam, Sadeh Norman, and Smith Noah, "Toward abstractive summarization using semantic representations", In Human Language Technologies: Annual Conference of the North American Chapter of the ACL, pp. 1077-1086, June-5, 2015.
[21] J. Flanigan, S. Thomson, J. Carbonell, C. Dyer, and N. A. Smith, "A discriminative graph-based parser for the abstract meaning representation." In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2014.
[22] Dohare, Karnick and Gupta, "Text Summarization using Abstractive Meaning Representation", In arvix, 17 July 2017.