20th International Conference on Advances in ICT for Emerging Regions (ICTer 2020): 300 – 301

Abstractive Web News Summarization Using Knowledge Graphs

M.V.P.T.Lakshika
University of Colombo School of Computing
Colombo, Sri Lanka
tlv@ucsc.cmb.ac.lk

H.A.Caldera
University of Colombo School of Computing
Colombo, Sri Lanka
hac@ucsc.cmb.ac.lk

W.V.Welgama
University of Colombo School of Computing
Colombo, Sri Lanka
wvw@ucsc.cmb.ac.lk

2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer) | 978-1-7281-8655-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICTer51097.2020.9325453

***Abstract*— The rapid progresses in digital data acquisition techniques have led huge volume of news data available in the news websites. Most of such digital news collections lack summaries. Due to that, online newspaper readers are overloaded with lengthy text documents. Also, it is a tedious task for human beings to generate an abstract for a news event manually since it requires a rigorous analysis of the news documents. An achievable solution to this problem is condensing the digital news collections and take out only the essence in the form of an automatically generated summary which allows readers to make effective decision in less time. The graph based algorithms for text summarization have been proven to be very successful over other methods for producing multi document summaries. The summary generated from knowledge graphs is more in line with human reading habits and possesses the logic of human reasoning. Due to the fast growing need of retrieving information in abstract form, we are proposing a novel approach for abstractive news summarization using the knowledge graphs to fulfill the need of having more accurate automatic abstractive text summarization.**

Keywords—News Documents, Abstractive Summarization, Knowledge Graphs, Data Mining, Natural Language Processing

I. INTRODUCTION

Within the recent years, we have seen extraordinary interest in news aggregation and browsing. Hence news websites becoming increasingly popular. Usually the Internet provides excessive quantity of information for every news topic than what is needed. Most of such electronic news collections lack summaries. Hence online newspaper readers are overloaded with lengthy text documents where smaller version would do. Extracting knowledge manually from a collection of newspapers may cause confusions, miss the track for user and time consuming. Due to this, the selection of the best collection of information for a particular news topic in a minimum conceivable time is still a challenge. Most of the existing applications for text summarization are extractive and only few applications are generate abstractive summaries. The graph based algorithms for abstractive text summarization have been proven to be very successful over other methods for producing multi document summaries. The summary generated from knowledge graphs is more in line with human reading habits and possesses the logic of human reasoning. Therefore, we propose a feasible solution to this problem by assembling the key information of a news event published in news websites and taking out only the essence in the form of an automatically generated summary which allows newsreaders to make effective decision in less time.

II. LITERATURE REVIEW

Automatic text summarization techniques such as cluster based, template based, ontology based, semantic graph based methods, machine learning, fuzzy logic, neural network approaches proven to be very useful over 50 years up to now [1],[2]. Automatic text summarization approaches are two fold as extractive and abstractive summarization [1]–[4]. Extractive summarization techniques extract important sentences or phrases from the original text documents based on the weight of statistical and linguistic features and produce a summary without changing the original text [1]. Abstractive summarization generates the summary by understanding the source text using linguistic methods and generate new sentences by improving the focus of a summary while reducing the redundancy rate [1], [2], [5]. Abstractive summarization is more complex because it requires deeper analysis of source document(s) and concept-to-text generation [4]. These techniques can be further classified as structured and semantic based approaches [2], [5]. Semantic based approaches provide more coherent, information rich and well-structured abstractive summary than structured based approaches. Due to the semantic representation, semantic graph based approaches have been proven to be very successful over other methods for producing multi document abstractive summaries[6]. According to the Gupta et al. [7] the semantic-based methods are not able to achieve similar performance to deep learning approaches. Machine learning approaches to generate automatic summaries [7]-[10] have greatly improved compared to the primitive text summarization methods, but they cannot be combined with the background information to obtain higher level abstraction [10] and cannot scale to the requirement of large data sets. In order to understand a text well enough, we need some background knowledge. Wu et al. [10] and Kalpa et al. [11] inspired by those ideas and used knowledge graph to generate automatic text summaries. Knowledge graphs capture domain specifics and add rich and explicit semantics to infer additional knowledge and encapsulate a large amount of knowledge for human and machine consumption [11].

Graph based approaches use ranking functions comprising of one or more sentences weighting features [12],[13] to determine the relevancy of each edge for inclusion into the summary. According to the literature, we have identified that traditional word co-occurrence measures like TF-IDF gives more importance to the words that are more frequent in the document. Hence, a pure ranking algorithm can be less effective because ranking alone cannot determine what kinds of features are selected for the summary. But, increasing the set of attributes to include semantic properties and topological graph properties in the ranking function yields statistically significant improvement for the abstractive summaries [12], [13]. Knowledge graphs lead to information overload, and hence proper summarization techniques need to be explored [11]. Techniques in data mining are widely used to explore data from various perspectives and uncovering unexpected relationships between pieces of text and summarizing it into useful information. Therefore, computers use text mining to discover valuable and interesting information and knowledge techniques. Among these techniques, Association Rule from unstructured textual data in large volumes using different Mining (ARM) can be used in discovering the relationships

978-1-7281-8655-9/20/$31.00 ©2020 IEEE
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on July 19,2023 at 08:30:24 UTC from IEEE Xplore. Restrictions apply.

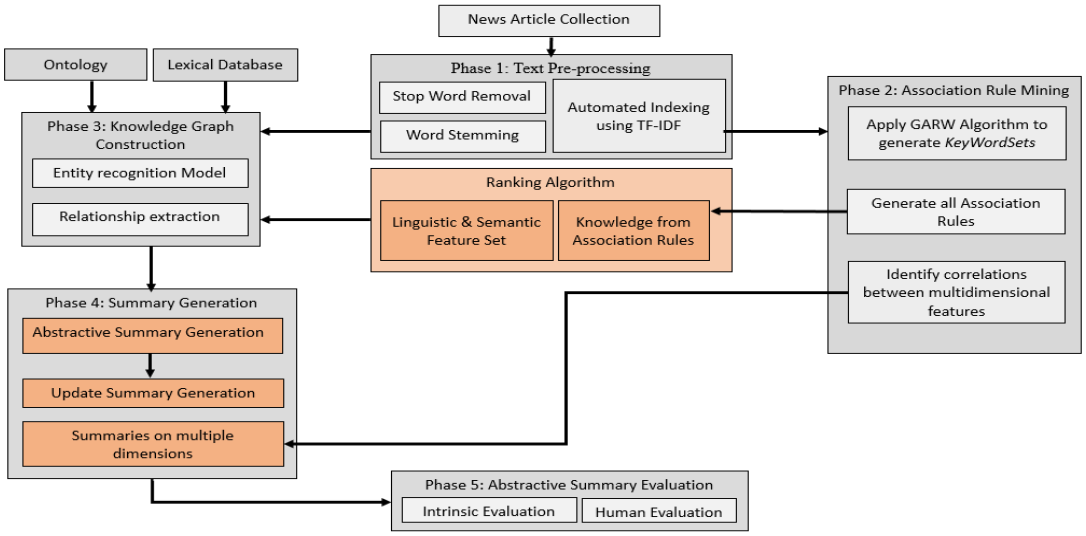

Figure 1: Proposed Methodology

for decision making process and discovering correlations between multidimensional features in the text documents [14], [15].

III. RESEARCH PURPOSE

We are proposing a novel approach to bridge the knowledge gap between Natural Language Processing and Data Mining fields to generate more cohesive, readable abstractive web news summarization using the knowledge graphs. This approach overcome the downsides in existing abstractive summary generation and enrich the accuracy of sentence ranking function using the knowledge derived from association rules in data mining for generating a better abstractive summary. According to the literature, any existing applications do not generate abstractive summaries on multiple dimensions (topics) in news articles and also do not generate 'Updated news summaries' for an existing abstractive summaries. The proposed approach generate both abstractive summaries on multiple dimensions or topics and update summaries to help readers to read and track news updates very easily.

IV. PROPOSED METHODOLOGY

As shown in the figure 1, after the text pre-processing steps, ARM is applied to the automated indexes generated from phase 1, and derived all the significant association rules from the phase 2. In the phase 3, knowledge graph is formed with two layers. Domain specifics in the newspapers are captured in the data layer and semantic layer adds rich and explicit semantics on top of the data layer to infer additional knowledge using ontologies and lexical database such as WordNet. A novel ranking algorithm which includes linguistic and semantic features in the news documents along with the knowledge mined from association rules in the phase 2 will be used to select the top-k entities from the knowledge graph. In the phase 4, abstractive summaries are generated using the SimpleNLG language generator. Updated abstractive summaries are generated for an existing abstractive summary. Generated correlations between multidimensional features in the news documents are used to generate multidimensional abstractive summaries. The collection of news documents released by Document Understanding Conference (DUC) will be used as the experimental dataset. In the phase 5, human evaluation using the domain experts will be carried along with the intrinsic evaluation techniques such as recall, precision, F1 measure, Pyramid and ROUGE evaluation.

REFERENCE

[1] R. Chettri, "Automatic Text Summarization," Int. J. Comput. Appl., vol. 161, no. 1, p. 3, 2017.

[2] N. Bhatia and A. Jaiswal, "Trends in Extractive and Abstractive Techniques in Text Summarization," 2015, doi: 10.5120/20559-2947.

[3] A. Patil, "Automatic Text Summarization," Int. J. Comput. Appl., vol. 109, no. 17, p. 2, 2015.

[4] N. R. Kasture, N. Yargal, N. Singh, N. N. Kulkarni, V. Mathur, "A Survey on Methods of Abstractive Text Summarization," 2014.

[5] D. K. Gaikwad and C. N. Mahender, "A Review Paper on Text Summarization," vol. 5, no. 3, p. 7, 2016.

[6] C. S. Yadav, A. Sharan, and M. L. Joshi, "Semantic graph based approach for text mining," in 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2014, pp. 596–601, doi: 10.1109/ICICICT.2014.6781348.

[7] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," Expert Syst. Appl., vol. 121, pp. 49–65, May 2019, doi: 10.1016/j.eswa.2018.12.011.

[8] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive Text Summarization Based on Deep Learning and Semantic Content Generalization," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 5082–5092, doi: 10.18653/v1/P19-1501.

[9] Z. H. Ali, "Multi-Document Text Summarization using Fuzzy Logic and Association Rule Mining", University of Technology – Department of Computer Science, no. 41, p. 19, 2018.

[10] P. Wu, Q. Zhou, Z. Lei, W. Qiu, and X. Li, "Template Oriented Text Summarization via Knowledge Graph," in 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, Jul. 2018, pp. 79–83, doi: 10.1109/ICALIP.2018.8455241.

[11] K. Gunaratna, "Semantics-based Summarization of Entities in Knowledge Graphs," p. 146, 2017.

[12] M. Subramaniam and V. Dalal, "Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method.," vol. 02, no. 02, p. 4, 2015.

[13] I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization," in 2012 Seventh International Conference on Computer Engineering & Systems (ICCES), Cairo, Egypt, 2012, pp. 132–138, doi: 10.1109/ICCES.2012.6408498.

[14] C. E. Crangle, "Text Summarization in Data Mining," in Soft-Ware 2002: Computing in an Imperfect World, vol. 2311, D. Bustard, W. Liu, and R. Sterritt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 332–347.

[15] H. Mahgoub, D. Rösner, N. Ismail, and F. Torkey, "A Text Mining Technique Using Association Rules Extraction," vol. 2, no. 6, p. 8, 2008.

Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on July 19,2023 at 08:30:24 UTC from IEEE Xplore. Restrictions apply.