ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Survey on Exploring Different Methods for Generating Content from Title

Amrutha C[1], Jayasree M[2]

PG Student, Dept. of Computer Science & Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Assistant Professor, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: There have been various approaches towards generating content from title. Some of them are topical poetry generator, recipe generation, story generation, abstract generation and so on. A survey is being done about different techniques applied in various domain regarding the content generation. Automated abstract generation is chosen as case study.

Keywords: Sequence-to-sequence model, GRU, Natural Language Generation, Attentive Revision Gate

I. INTRODUCTION

What is an abstract? Why is it important? What is its significance in a scientific paper? A scientific paper abstract should always focus on the keywords specified in the title. A typical Recurrent Neural Network (RNN) based approach easily loses focus. So a title attention is introduced into a sequence- to-sequence model to guide the generation process so that the abstract is topically relevant. Today, machine learning, Artificial Intelligence (AI), and neural networks are developing rapidly. So it would be weird not to take advantage of this development by using an online abstract generator. An abstract is an important part of any academic or professional paper. These brief overviews serves as a summary of what the paper contains, so it should succinctly and accurately represent what the paper is about and what the reader can expect to find. The purpose of abstract in technical literature is to facilitate quick and accurate identification of the topic of published papers. The abstracter's product is always influenced by his/her background, attitude and disposition. To bring out the notable points of an author's argument calls for skill and experience. As a result, a considerable amount of qualified manpower that could be used to advantage in other ways should be diverted to the task of facilitating access to information. This ubiquitous problem is being aggravated by the ever increasing output of technical literature. But another problem perhaps equally acute is that of achieving consistence and objectiveness in abstracts. There are various techniques used for text generation, such as abstract generation [1] [2], poetry generation [3], recipe generation [4] and so on. The proposed framework, focus on generating abstract from the data set of 10,874 paper title abstract pairs from the ACL Anthology Network. The preparation of abstract is an intellectual effort, requiring general familiarity with the subject.

In the survey conducted, an automated routine writing system is explored with paper abstract writing as a case study. Two sets of Turing test is being done and the system provides efficient results. In addition, the dataset used for experiments mainly focus on NLP (Natural Language Processing) data and hence produce best result in that domain. Automatic abstract generation is quite useful when one is making an attempt to grasp the content of a paper. It serve as an assistive technology for human to write paper abstracts. When one is conducting a survey with a large number of scientific documents, it will be helpful for him/her to understand the idea of the paper by just reading the abstract. So if an automated abstract generation system is available that creates abstracts from title, instead of reading the whole paper and then identifying the key points it becomes really helpful and saves time. The task of abstract generation is very complex because the concepts contained in the titles are usually limited, so the learning space for the generator is huge, which hinders the quality of the generated abstract. As a result many difficulties or challenges are being faced in the process of abstract generation from title. Machines lack common sense knowledge and logical coherence. Human written abstracts are generally more specific, concise, and engaging. This paper is organized as follows: Section II gives a formal definition of the automatic abstract generation and an example. Section III discusses about various approaches used in different domains and provides a comparison between them. Section IV discusses various future research directions of abstract generation. Section V gives brief concluding comments.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

II. AUTOMATIC ABSTRACT GENERATION

Text generation task is one of the NLP related task which can be implemented by using deep learning models (particularly Recurrent neural network), artificial intelligence and so on. The system being studied [1] deals with the automatic generation of abstracts from the given title.

The architecture consists of two section: A writing network and an editing network. The writing network takes a title as input and generates the first draft. The editing network takes both the title and previous draft as input to iteratively proof-read, improve, and generate new versions. Writing Network is based on an attentive sequence-to-sequence model. The architecture is shown in the figure 1.

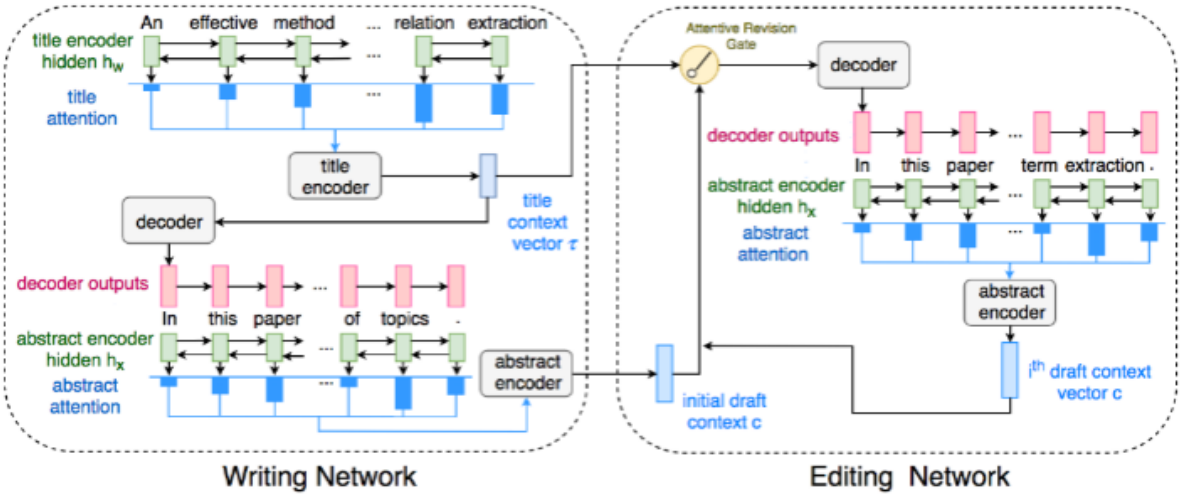

Fig. 1 Architecture of abstract generation system.[1]

The problem of abstract generation can be formally defined as: Given a title as input, the proposed system generates a concise and coherent abstract that is topically relevant.

The Writing Network is based on an attentive sequence-to- sequence model. A bi-directional gated recurrent unit (GRU) is used as an encoder, which takes a title as input and for each token, the encoder produces a hidden state. In this manner, the title is encoded. A soft-alignment attention mechanism is adopted to capture the correlation between the title, T and the abstract draft. It enables the decoder to focus on the most relevant words from the title.

A. Domain Based Approaches

The concepts contained within the titles are usually limited and hence the learning space for the generator is huge, which hinders the quality of the generated abstract. Compared to the title, the generated abstract contains more topically relevant concepts, and provide better guidance. Therefore, an Editing Network, is designed which, besides the title, takes the previously generated abstract as input and iteratively refines the generated abstract. According to the study conducted, it is found that an attempt regarding automatic abstract generation task dated back to 1958. Luhn et.al. [5] have done some exploratory research on automatic methods of obtaining abstracts. In this paper, instead of sampling at random, as a reader normally does when scanning, the new mechanical method selects those among all the sentences of an article that are the most representative of pertinent information. These key sentences are then enumerated to serve as clues for judging the character of the article. Thus, the citations of the author's own statements constitute the "auto-abstract".

The programs for making auto-abstracts must be based on properties of writing ascertained by analysis of specific types of literature. Because the use of abstracts is an established practice in science and technology, it seemed desirable to develop the method first for papers and then for the articles in this area. To determine which sentences of an article may best serve as the auto-abstract, a measure is required by which the information content of the entire sentences can be compared and graded. Since the suitability of each sentence is relative, a value can be assigned to each in accordance with the quality criteria of significance.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

The "significance" factor of a sentence is derived from an inspection of its words. The frequency of word occurrence in an article provide a useful measurement of word significance. Also the relative position within a sentence of words having given values of significance furnishes a useful measurement. The significance factor of a sentence is based on a combination of these two measurements. Emphasis is taken as an indicator of significance. This system does not propose to differentiate between word forms. Also no consideration is given to the meaning of words. Logical and semantic relationships are not considered.

Abstract generation is, like Machine translation, one of the ultimate goals of NLP. However, since conventional word-frequency based abstract generation system [5] are lacking inter- sentential or discourse structural analysis, they are liable to generate incoherent abstract. On the other hand, conventional knowledge or script based abstract generation system ([6], [7]) owe their success to the limitation of the domain, and cannot be applied to document with varied subjects.

Ono et.al. [8] proposed an automatic abstract generation system for Japanese expository writings wherein the system first extracts the rhetorical structure i.e. the compound of the rhetorical relations between sentences. Less important parts in the extracted structure is being removed to generate abstract of desired size. In this model, discourse structure is defined as rhetorical structure. The rhetorical structure represents relations between various chunks of sentences in the body of each section.

The system generates the abstract of each section of the document by examining its rhetorical structure. The technical papers contain so many rhetorical expressions in general as to be expository. That is, they render many linguistic clues and the system can extract the rhetorical structure exactly. Consequently, the structure can be reduced further and the length of the abstract gets shorter, without eliminating key sentences. The limitations of the system are mainly due to errors in the rhetorical structure analysis and the sentence selection type abstract generation.

Kaneko et. al. [2] proposes an effective approach to create abstracts for social scientific papers. Important keywords, readability as an abstract, and features of social scientific papers are taken into account for the generation. Most papers in social science domain are very long and some of them don't even have any abstracts at all. In this work, in order to perform textual analysis and importance degree estimation for words or phrases, following five lexicon-files are created. Adverb Lexicon, Sentence-End Expression Lexicon, Conjunction Transformation Lexicon, Indispensable-case Lexicon, Conjunction Lexicon.

Also, three lexicons specialized in social science domain is used. The first one is a called Keyword Dictionary, containing the words extracted from two sociological dictionaries. The second lexicon is the Dictionary. Five kinds of noun phrases from a social scientific literature database is being extracted.

Three main modules have been developed in this system to generate the abstract: sentence processing, importance degree estimation, and abstract generation. Experimental results have shown the effectiveness of this proposal in comparison with another existing summarization tool, especially when character-based calculation is used to estimate the necessary number of constituents for abstract generation. The major drawback of this system is "Pronouns are met too frequently", "Too many long sentences exist in the abstract".

Recurrent Neural Network (RNN) architectures have proven to be well suited for many Natural Language Generation (NLG) tasks. Previous neural generation models typically generate locally coherent language that is on topic; however, overall they can miss information that should have been introduced or introduce duplicated or superfluous content. These errors are particularly common in situations where there are multiple distinct sources of input or the length of the output text is sufficiently long.

Kiddon et.al. [4]proposes an idea called neural checklist model that generates globally coherent text by keeping track of what has been said and still needs to be said from a provided agenda. Existing RNN models may lose track of which ingredients have already been mentioned, especially during the generation of a long recipe with many ingredients. The two key features of this model are that it (1) predicts which agenda item is being referred to, if any, at each time step and (2) stores those predictions for use during generation. The power of the model is in generating long texts, but the experiment shows that this model can generalize well to other tasks with different kinds of agenda items and goals.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Automatic recipe generation is difficult mainly due to the length of recipes, the size of the vocabulary, and the variety of possible dishes having same ingredients. One of the major limitations of this model is the situation in which a number of recipes that have very similar text but make different dishes. So there arises a chance for error occurrence. Another drawback is the simple string match performed so that it miss certain referring expressions (e.g., meat to refer to pork).

The methodology proposed by Ghazvininejad et al. [3], introduces a new language model wherein the system generate poems given a user specified topic. Poems generated obey rhythmic and rhyme constraints. The system uses special techniques, such as rhyming word choice and encoder-decoder modeling, to keep the poem on topic.

One of the major drawback of this system is that the system generate poems of a particular type only. i.e, the iambic pentameter. It is one of the commonly used meters in English poetry. It is a rhyme scheme in which each sonnet line consists of ten syllables (unstressed syllable followed by stressed syllable). Since trained on song lyrics, language model tends to generate repeating words, like never ever ever ever ever. To solve this problem, a penalty is applied to those words. Another problem is that discourse structure problem may arise.

Sequence-to-sequence (seq2seq) neural networks have achieved state of the art performance on a variety of text generation tasks. Previous work on story generation has focused largely on using various content to inspire the stories such as photos, sequences of events. The technology proposed by Fan et.al [9] is a neural network based approach where a premise is given as input. The system generates a story as output.

It uses CNN (Convolution Neural Network) which is a new approach in this text generation task. Since premise is given as input, the generated story remains consistent and also has structure at a level beyond single phrases. Many seq2seq RNN architectures have been used for story generation. But it is found that neural based techniques are better than seq2seq approaches. Since this technology is introduced recently, the dataset used have some deficiencies which could be further improved. If a good dataset is used, it increases the accuracy further.

Lebret et.al [10] introduces a neural model for concept to text generation that scales to large, rich domains. The system generates biographical sentences from fact tables on a new dataset of biographies from Wikipedia. From the experiments it is found to generate fluent descriptions of arbitrary people based on structured data. In this paper, they have only focused on generating the first sentence.

III. FUTURE RESEARCH DIRECTIONS

Various techniques have been applied in the text generation task. The method adopted by [1] is the recent one. Even though it generate topically relevant abstract, the size of the abstract is not considered. Since the constraint size is not specified, the generated abstracts will not be uniform.

Most of the abstract generation task is domain dependent ([2], [1]). The system proposed by [2] uses social science as the domain and [1] uses NLP as domain. The system performs well by using domain dependent dataset. So using a number of dataset of different domain it is possible to model a system that is domain independent.

CONCLUSION

Several techniques have been proposed by different authors regarding automatic text generation task. The attempt to generate abstract has begun from late 19th century. Initial approach was based on simple word frequency based abstract generation system. As long as new inventions or discoveries are made, it is essential to document it for future reference. Drafting a scientific paper is a tedious task. One of the major hurdles is to create the abstract which summarize the whole content in a gist. Here comes the importance of an automatic abstract generation system. Here, a new paper abstract generation system which generates a concise and coherent abstract with the help of title is being surveyed. Several experiments have been conducted in the literature, and it is found that generate topically relevant abstracts can be generated using writing editing mechanism.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

REFERENCES

[1]. Q. Wang, Z. Zhou, L. Huang, S. Whitehead, B. Zhang, H. Ji, and K. Knight, "Paper abstract writing through editing mechanism," arXiv preprint arXiv:1805.06064, 2018.
[2]. M. Kaneko and D. Han, "An abstract generation system for social scientific papers," in Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27), pp. 57–65, 2013.
[3]. M. Ghazvininejad, X. Shi, Y. Choi, and K. Knight, "Generating topical poetry," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1183–1191, 2016.
[4]. C. Kiddon, L. Zettlemoyer, and Y. Choi, "Globally coherent text generation with neural checklist models," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 329–339, 2016.
[5]. H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of research and development, vol. 2, no. 2, pp. 159–165, 1958.
[6]. D. Fum, G. Guida, and C. Tasso, "Tailoring importance evaluation to reader's goals: a contribution to descriptive text summarization," in Proceedings of the 11th coference on Computational linguistics, pp. 256–259, Association for Computational Linguistics, 1986.
[7]. W. G. Lehnert, "Plot units and narrative summarization," Cognitive Science, vol. 5, no. 4, pp. 293–331, 1981.
[8]. K. Ono, K. Sumita, and S. Miike, "Abstract generation based on rhetorical structure extraction," in Proceedings of the 15th conference on Computational linguistics-Volume 1, pp. 344–348, Association for Computational Linguistics, 1994.
[9]. A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," arXiv preprint arXiv:1805.04833, 2018.
[10]. R. Lebret, D. Grangier, and M. Auli, "Neural text generation from structured data with application to the biography domain," arXiv preprint arXiv:1603.07771, 2016.
[11]. L. Wang and W. Ling, "Neural network-based abstract generation for opinions and arguments," arXiv preprint arXiv:1606.02785, 2016.
[12]. E. Greene, T. Bodrumlu, and K. Knight, "Automatic analysis of rhythmic poetry with applications to generation and translation," in Proceedings of the 2010 conference on empirical methods in natural language processing, pp. 524–533, Association for Computational Linguistics, 2010.
[13]. J. Zhang, Y. Feng, D. Wang, Y. Wang, A. Abel, S. Zhang, and A. Zhang, "Flexible and creative chinese poetry generation using neural memory," arXiv preprint arXiv:1705.03773, 2017.
[14]. J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," arXiv preprint arXiv:1603.06393, 2016.
[15]. A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," arXiv preprint arXiv:1704.04368, 2017.
[16]. T. Liu, K. Wang, L. Sha, B. Chang, and Z. Sui, "Table-to-text generation by structure-aware seq2seq learning," arXiv preprint arXiv:1711.09724, 2017.
[17]. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proceedings of the 26th annual international conference on machine learning, pp. 41–48, ACM, 2009.
[18]. R. Koncel-Kedziorski, I. Konstas, L. Zettlemoyer, and H. Hajishirzi, "A theme-rewriting approach for generating algebra word problems," arXiv preprint arXiv:1610.06210, 2016.
[19]. O. Polozov, A. M. Smith, L. Zettlemoyer, S. Gulwani, and Z. Popovic, "Personalized mathematical word problem generation.,"