



e-ISSN: 2319-8753 | p-ISSN: 2320-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.512



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

A Content Transfer Approach: Rule based Table to Text Generation

Divya Visakh¹, Dr. Reena Murali², Dr. Ramesh Babu K R³

P.G. Student, Department of Computer Science & Engineering, Government Engineering College, Palakkad, India¹

Professor, Department of MCA, RIT, Kottayam, India²

Associate Professor, Department of Information Technology, Government Engineering College, Palakkad, India³

ABSTRACT: Natural language generation is considered as the sub-field of natural language processing. Natural language generation is a process of producing text from the data provided as input to the system. The content transfer is the process of transferring the contents from one source to another. Here the content transfer focuses mainly on the structured form of data (i.e. Table) to the set of sentences (i.e. Sentences). The tables used in this system are the Wikipedia Infobox. Wikipedia Info boxes are considered to be a subset of information that is used to represent the knowledge of the particular subject in a nutshell. Info boxes are the factual tables with a set of field-value pairs. This system is domain dependent as it focuses only on the famous Malayalam language author's infoboxes. Firstly the system is set up in such a way to extract or scrape the infoboxes from the Wikipedia website. After extracting the required field of information, it is converted into the CSV format. In order to generate the biography, it is very important to know the gender of the author. With the help of Naive Bayes, the system identifies the gender of the author whether it is female or male. Then a certain different sets of rules are written to create the corresponding sentences for each field in the Infobox. After this module, the results from the rule-based phase are ordered into a set of sentences that take up the shape of the author's biography. The challenges include the ordering of sentences and dataset creation. The dataset includes 197 well-known Malayalam author's information taken from their corresponding Wikipedia infoboxes. In order to analyze the system results, various reading rates are calculated. Along with that, a manual survey was conducted on the results, 90% accuracy is obtained for the results.

KEYWORDS: Text Generation, Natural language processing, Rule-based methods, Biography, Wikipedia Infobox and Malayalam Literature.

I. INTRODUCTION

In this fast paced internet era, there are wide varieties of information that are published online each day as a source of information to the world. Although there are different websites to share the information, the most trusted and high ranked list while searching a query will be Wikipedia. Wikipedia is considered to be the most trusted free website that serves information on wide variety of information for different genres. In the Wikipedia website, infoboxes plays an important role. Before reading the whole description of one famous person, user always looks upon the infoboxes which consists of essential features that helps to describe the person completely.

Natural language text generation [1] is considered to sub portion of the field of natural language processing. Text generation aims to generate text from the computer data. It converts the data in any form to natural language representation. The text will be generated on the basis of the information that is the collected data and the functionality of the task that is provided by the user. Text generation [2] mainly focuses on the content and structure. Both are very relevant to generate text. A proper and sound content along with the proper structure only will work in a perfect harmony to generate the accurate information in the text format.

Content transfer is the process of effective transmission of contents from one source to another. Every day there is lot of content transfer taking place, a proper grounded way of content transfer is implementing in this system. Natural language descriptions from structured data [3] like tables are required in various fields of applications such as generating descriptions of products, authors, food, furniture, companies etc. A table usually comprises in the combination of field and value pairs. The field represents a property of the entity (e.g. name) and the value is a set of possible assignments to this property (e.g. name = divya). The Wikipedia [3] info-box describes as a table of facts with different fields and values. In order to generate a sentence based on the Infobox, the system should include the relevant

fields for better knowledge. Table-to-text generation always follows the steps of NLG. That is including the content planning for selecting specific content from the input and determines the structure of the output text that a user needs. Next is sentence planning [4] that determines the structure and lexical content of each sentence and is produced during the process for the effective knowledge transfer. Last one is surface realization which converts the sentence plan to a proper set of ordered sequence sentences. A sample system approach is present in the figure 1.

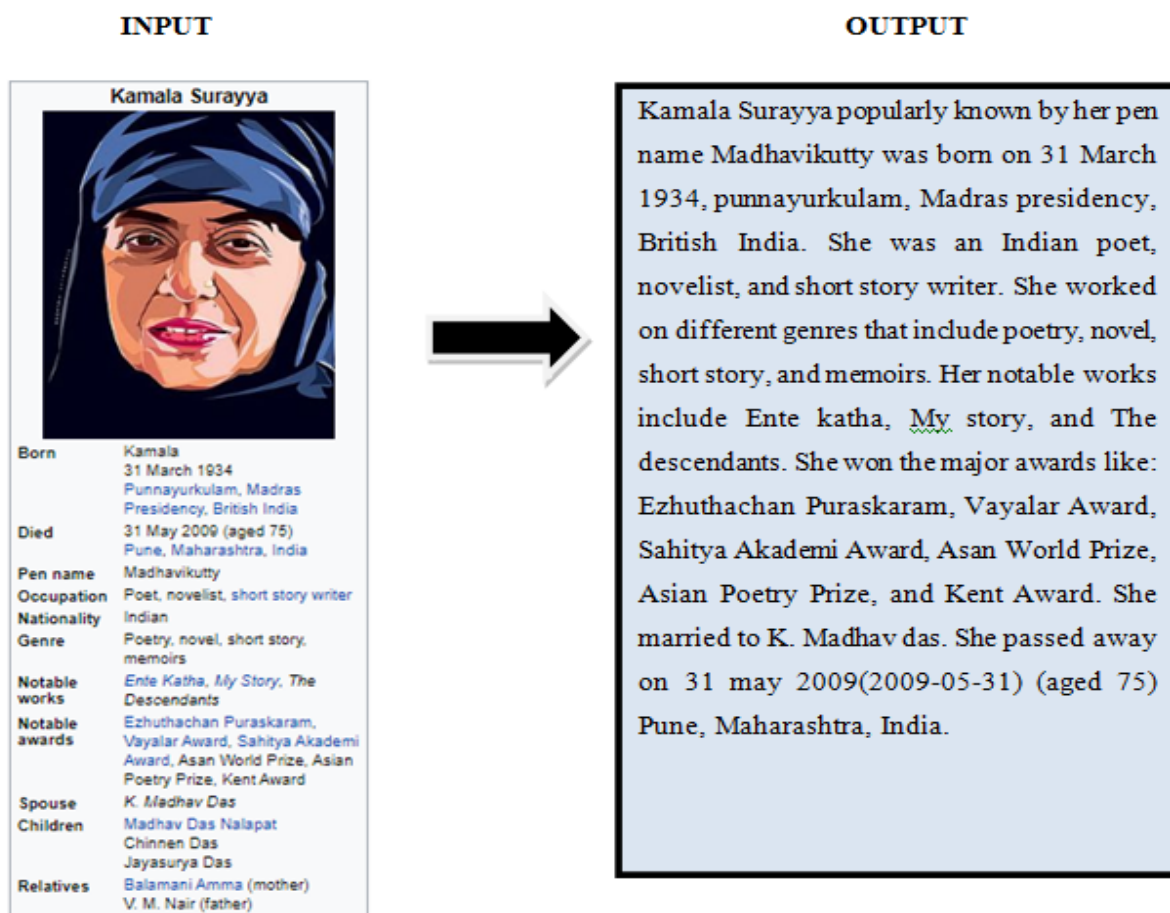


Figure 1. Sample System Output Approach

In order to understand a table and to generate the most meaningful, relevant and sentences, a table-to-text model always faces different challenges. Below describes the different challenges that the system faced while generating the text.

- Which records in the table should be included in the introduction and how to arrange the order of these records before wording?
- Which words or phrases in the table should be more focused on to paraphrase?
- Which words or phrases in the table should be discarded while generating text?
- How to identify the gender of each author through the name?

The paper is organized as follows: Section I contains the introduction of abusive language detection system. Section II contain the related work of abusive language detection; Section III discusses about the system architecture, deep learning model and dataset used in this work; Section IV presents the implementation phases, results, performance analysis and performance comparison of proposed work with existing techniques. Section V concludes research work with future directions.

II. RELATED WORK

This section deals with the previous research works for table to text generation.

Creating writings from organized information [5] (e.g., a table) is significant for different common language preparing errands, for example, question noting and discourse frameworks. Analysts analyzed the use of different neural language models and encoder-decoder frameworks for table-to-text generation. In any case, these neural networks don't organize the sentences properly. At the point when a human composes a synopsis in view of a given table, the person in question would presumably consider the content request before wording. The text generation process is based upon the RNN with attention to table contents. This model focused on the order of contents by a link matrix, based on computing a link-based attention. Then a self-adaptive gate neural network adjusts the content- and link-based attention mechanisms.

Table-to-content generation [6] intends to produce a depiction for a true table which can be seen as a lot of attribute-value records. To encode both the substance and the structure of a table, a novel structure-mindful seq2seq design is proposed which comprises of field-gating encoder and portrayal generator with double consideration. In the encoding stage, it updates the cell memory of the LSTM unit by a field gateway and its comparing field an incentive so as to join field data into table portrayal. In the translating stage, double consideration system which contains word level consideration furthermore, field level consideration is proposed to demonstrate the semantic significance between the created text and the table.

A fused bifocal attention mechanism [7] which combines both micro and macro level information is introduced. It included a gated orthogonalization mechanism to ensure that a field to know whether it is remembered for a few time steps and then forgotten. Dataset contains fact tables about people and their corresponding one line biographical descriptions in English. In addition they also introduced two similar datasets for French and German.

A neural network (NN) architecture [8] which incorporates two techniques: content selection and planning without sacrificing the process of end-to-end training. It decomposed the generation task into two different stages. With the help of the corpus of data records which are already paired with descriptive documents. Then first generate a content plan highlighting which information should be mentioned first and the order of arrangement. Then only the document is generated by taking the content plan into account. It worked on a ROTOWIRE dataset.

TOTTO is a huge English table-to-content dataset [9] that presents both a controlled age task and an information explanation process in view of iterative sentence correction. It gave a few best in class baselines, and exhibited TOTTO could be a helpful dataset for displaying research just as for creating assessment measurements that can all the more likely distinguish model better performances.

A table-to-successive (Table2Seq) [10] neural system model, accepts a table as information and produces a characteristic language sentence identified with that table. The Table2Seq model is in an encoder-decoder structure, which has been tried to be viable and generally applied in numerous errands. The Table2Seq design incorporates an encoder and a decoder. The encoder first speaks to a table as consistent vectors with regards to both the content and the structure data. At that point, through completely utilizing the table portrayal, the decoder creates a sentence word by word. During the deciphering methodology, a duplicating system is received making the model able to do imitating uncommon words from table cells, properties, and subtitles. The upside of such a neural generation model is that it can be prepared in a start to finish style with back-propagation.

Reports [11] are consequently created by amassing a majority of provisos chose from a library of statements put away in a computer framework. A standard set is doled out to every one of the conditions. Each standard set gives at any rate one principle that must be fulfilled so as to incorporate the proviso related therewith in an archive. After report parameters are gone into the PC framework, each standard set is tried to distinguish those that are fulfilled by the record parameters. The conditions to which the distinguished principle sets are doled out are recovered and gathered into the report. The record age framework is especially reasonable for the production of protection arrangement reports. At the point when important, protection approach provisions are supplanted with support conditions dependent on underwriting choices made by a client. A support determination list is produced by testing rule sets related with the supports.

III. METHODOLOGY

This section gives the system architecture and the overall implementation details of the proposed system. How well the proposed text generation system works on the Malayalam wikipedia infobox dataset is also explained.

A. Dataset :

Dataset is fully generated as a part of the project to focus upon the Malayalam Language authors. Instead of millions numbers of various famous person as dataset, the system fully focus on the Indian Malayalam Authors. A total of 197 well known authors infoboxes are extracted from the Wikipedia website through scraping and storage in csv format. The csv file contains two columns. One column represents the fields and the other column represents the corresponding values of the field. A 100% successful extraction of data is obtained in the dataset creation stage without any losses.

Info boxes include the salient key features to represent a particular person/author. This information is fully about the name, place, date of birth and death, occupation, genres, pen name, family background, famous works and notable awards. From this nutshell information, a biography is generated. Out of 248 author only 197 authors have valid infoboxes with necessary information that help us to generate a useful biography. Remaining infoboxes are sometimes empty or if present only two or three fields present with no information.

The dataset consists of:

- 197 author's infoboxes.
- More than 10 different field information.

B. System Architecture:

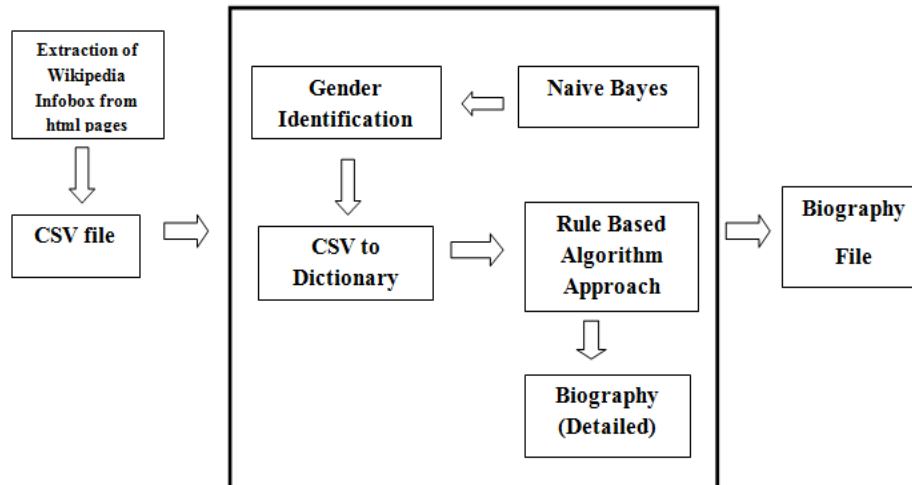


Figure 2. Basic architecture for Table to Text generation

The proposed system creates relevant set of statements that compromises to form the corresponding biography of an author. It includes all the necessary facts that help to describe the person's life efficiently.

Figure 2 shows the basic architecture of the proposed table to text generation system. The overall system architecture compromises of five different modules or phases.

1. Dataset Creation Module
2. Gender Identification Module
3. Dataset dictionary conversion Module
4. Rule based approach Module
5. Biography Generation Module

I. Dataset Creation Module :-

The first module is really an important part in the whole system as it focuses only on the famous Malayalam literature authors. So there is no proper dataset that provides the information of the Malayalam authors in the form of infoboxes. The text based biographies can be always represented as sequences of facts and events that develop to describe a person's life. The facts include different aspects related to personal characteristics like date and place of birth, death date, education, career, occupation, famous works, notable awards and family background [12].

Infobox is a well-formed table that gives a short and precise description of important facts related to the person. These are the different fields that this system focuses upon.

- Name: Name of the author.
- Date of Birth & Date of Death: Extracted as context phrases such as 'born on', 'birth', 'died', etc.
- Place of birth: Identify the place of the birth.
- Pen name
- Famous works
- Genres
- Occupation
- Notable awards

Dataset includes about 197 author details which is in csv format. The different steps that happened in this module is explained below:

- Wikipedia Info boxes are scraped from the free and trusted website of Wikipedia. The web scraping is done with the help of BeautifulSoup library. It helps to strip a part of information with the help of corresponding tag elements that models the whole website.
- Secondly as the infoboxes represents as the combination of field-value pairs, a standard and systematic way of storage of data is needed. So the data retrieved is transferred into csv format.
- Inside the csv file, two columns are present; First one is to represent the various fields like name, date of birth. The second column provides the corresponding values of each field.

II. Gender Identification Module :-

In this module, it fully focuses on the determination of gender of the author his/her full name. The gender of the author will not be given on infoboxes. Identifying the gender is very much important to generate biographies. Biographies never repeat the name of the person more than once. Instead it uses pronouns to describe the person. So gender identification is necessary.

Naive Bayes works as a solution by an easy yet efficient way of gender identification. It is based on understanding the pattern of the whole data set to build feature set on which model will be trained. As per various NLTK documentation research, it is observed that female names most likely to end with 'a', 'e', 'i', 'y' whereas male names are with 'k', 'r', 'o', 's', 't', 'h', 'm', 'n'. Here the system used the names corpus that has 2 categories male name & female name. Then created a methodology to extract last character of the name & this will help to generate feature set based on the assumption. The assumption is that the last character can be most useful attribute to identify gender of the name. After importing the required modules to collect the names for female/male. Then list of tuples are created with name & its gender for female & male name list also combined Female & Male List of tuples followed by shuffling it. By splitting the data into two sets which is train/test data set. Thus this learning model helped to identify the gender based on the most likelihood probability.

III. Dataset dictionary conversion Module :-

The first step in this module is to clean the unnecessary information that is extracted from the infoboxes. Cleaning the csv is very important to generate accurate set of sentences. With the help of nltk, cleaning of the data is done.

Next step is conversion of csv dataset into dictionary format. A dictionary is considered as a collection of changeable and indexed data values. Dictionaries are always written within curly brackets. Dictionary holds a key: value pair

where Key value is provided in the dictionary to make it more optimized. Every key corresponds to some values. Here in this module, dictionary conversion helps to provide a systematic data as input for the next important module.

IV. Rule based approach Module :-

Rule-based system (RB) is widely known as knowledge-based systems (KB). A rule-based system performs actions or operations based on some preset rules and conditions. It is a straight-forward, easy to implement approach and can be integrated into different systems. Another important feature is that it can deliver fast-paced as well as instant solutions for the various user query handling models. The rule-based system [11] always tries to work just like a human expert knowledge for the NLP applications. In the RB systems, the information or data from the corpus is fed into the system; all the rules which are developed will be applied on the input data and results on a output that satisfies the system goal. Rules-based systems are a simple kind of artificial intelligence, which use a series of IF-THEN statements that guide a computer to reach a conclusion or recommendation. Figure 3 gives basic ideas of the working of rule-based system.

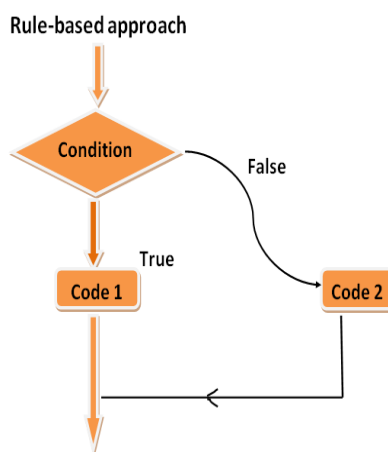


Figure 3: Rule based system

Two Key Components of rules-based system [13] includes first a set of facts that describes or supports a situation and secondly, a set of rules that are used on how to deal with those facts. The set of facts are known as the knowledge base. Facts are a combination of the data i.e name, place of birth etc. The set of rules are known as the rules engine. Rule engine focuses on the rules that describe the relationship between the IF and the THEN statements. From these two basic concepts, it is possible to build a basic system to generate a text in a particular or pattern. In this system, 8 different rules are written to generate a particular of sentence. The rules helped to combine the goal and transfer the content in a systematic and sequential order. The Rule-based systems are able to offer quicker, solution to the problem and are really cost-efficient. Other important features include availability, cost efficient, high accuracy and less error rate and steady response.

V. Biography Generation Module :-

Biography [14] is considered as the description to represent a person's life. The event includes all the basic information regarding that person. After retrieving the different key features of information, a set of sentences are generated into list. These sentences [15] are then again ordered into a systematic structured format. These sentences itself form a story that covers all the important points of that person. Biography of the author that is generated as an output from the previous phase was written to a text format.

IV. EXPERIMENTAL RESULTS

A. Tools and Libraries

The tools and libraries used in the proposed system are given in below:

Table I: Tools used

Task	Method Used
Infobox scraping from website	BeautifulSoup Library
Csv file data conversion	Pandas and Csv Library
Dataset cleaning	NLTK
Naive-Bayes Approach	NLTK, GaussianNB library
Rule-based Approach	NLTK, Rule-Engine library

B. Results

An example of infobox and its corresponding biography generation is as given below:


	
Born	Madath Thekkepaattu Vasudevan Nair . 9 August 1933 (age 86) ^[1] Kudaliur, Ponnani taluk, Malabar District, Madras Presidency, British India (present-day Pattambi taluk Palakkad district, Kerala, India)
Occupation	Novelist, short story writer, screenplay writer, film director
Language	Malayalam
Residence	Sithara, Calicut, India ^[2]
Alma mater	Victoria College, Palakkad
Genre	Novel, short story, children's literature, travelogue, essays
Subject	Social aspects, Oriented on the basic Kerala family and cultures
Literary movement	Realism Fyodor Dostoevsky Anton Chekhov Guy de Maupassant Vailom Muhammed Basheer
Notable works	<i>Neelukettu</i> <i>Randamoozham</i> <i>Manju</i> <i>Kaalam</i> <i>Asuravithu</i> <i>Irutinte Athimavu</i>
Notable awards	Padma Bhushan, Njanapeedam, Kendra Sahitya Akademi Award, Kerala Sahitya Akademi Award
Spouse	Prameela (1965–1976; divorced) Kalamandalam Saraswathy

Figure 4: Author Infobox

Result obtained from the system is as follows ;

"M T Vasudevan nair was born on (1933-08-09) 9 august 1933 (age 86) kudallur, ponnani taluk, malabar district, madras presidency, british india (present-day pattambi taluk palakkad district, kerala, India). He is a novelist, short story writer, screenplay writer, and film director. He worked on different genres that include "novel, short story, children's literature, travelogue, and essays". His notable works include Naalukettu, Randamoozham, Manjukaalam, Asuravithu, and Iruttinte athmavu. He won the major awards like : Padma bhushan, Njanapeedam, Kendra sahitya akademi award, and Kerala sahitya akademi award. He married to Pameela(1965-1976; divorced), Kalamandalam saraswathy(1977 to present)."

C. Performance Analysis

As the system fully relies on the fundamental natural language processing techniques, it is difficult to apply the popular sequence to sequence text result analysis technique like BLEU. The continuous study implies that even though it is developed for the machine translation evaluation, it has several drawbacks when worked with machine and neural network based system. BLEU doesn't fully handle morphological rich language and doesn't give importance to the sentence structure. It doesn't match with human judgments. So a proper and stable analysis is needed to analyze the system. Here the system mainly analyzed on two different aspects. They are:

1. Readability Score
2. Manual Response Analysis

1. Readability Score:

Readability is considered to know the ease with which a reader can understand a generated text. In natural language processing, the readability of text always depends on its content (the complexity of its vocabulary and syntax). It fully focuses on the chosen words, and how they organised into sentences and paragraphs for the readers to comprehend easily. The main objective in generating text system is to pass along genuine information so that the reader think is worthwhile as it provides knowledge. If the system fails to convey that information, efforts are wasted. In order to engage the reader to read, it's critical to present information to them clearly and precisely. It assists the system in improving the text to make it more understandable and can engage a larger audience.

A readability score is considered to be a numeric value that indicates how difficult (or easy) it is to read and understand a text by the readers. There are different tests to determine the readability and the standard of the information it delivers. Readability is not only considered to be the only method for determining readability but also as the predictions of reading ease or accuracy of the generated text. There are different tests. Figure 5.8 and 5.9 shows the scores obtained from different level score methods. Some of them that are utilized in this system are :

- Flesch-Kincaid Grade Level
- Flesch Reading Ease

2. Manual Response Analysis:

This analysis depends upon the response obtained from different people's perspective. A human analysis helps to check how much the system reached the level of human understanding. For this, a Google feedback form was created with 7 randomly chosen generated outputs of famous Malayalam authors. For every generated output questions, three options were assigned. The three options were internally given three values too. So the people assigned different options for the 7 questions based on the readability, meaning and understanding. Based on these results, accuracy is calculated for the system.

After obtaining the responses, the calculation is done to find the accuracy. First all the option 1-responses obtained are combined together for each question. The same method was followed for the other 2 options. Then with the help of assigned score for each option, a score is obtained for each option. Finally the whole score is divided by the 60 responses obtained and converted to the respective percentage- i.e.: Accuracy.

Table II. Scores assigned for 3 options

Options	Score Assigned
Perfect- Readable	1
Average	0.75
Need Improvements	0.5

Table III. Response percentage received for the 7 sample outputs

Question No	Option - 1	Option - 2	Option - 3
1	75%	21.7%	3.3%
2	81.7%	13.3%	5%
3	55%	25%	20%
4	73.3%	16.7%	10%
5	83.3%	8.3%	8.3%
6	65%	20%	15%
7	71.7%	25%	3.3%

Table IV. Response numbers received for the 7 sample outputs

Question No	Option - 1	Option - 2	Option - 3
1	45	13	2
2	49	8	3
3	33	15	12
4	44	10	6
5	50	5	5
6	39	12	9
7	43	15	2

Table V. Score obtained for each option from the 7 samples

Option No:	Total Score
1	298
2	58.5
3	19.5

D. Overall Performance

The overall performance of this proposed system is given in Table VI.

Table VI. Overall Performance Analysis

Analysis	Score
Manual Performance Analysis	90%
Reading score	College grade level

V. CONCLUSION

The importance of Natural language text generation is increasing in the fast paced internet era. Text generation helps in various fields effectively like Conversational AI, Question answering, Report generation etc. This domain dependent system fully aims on developing a simple yet powerful text generation mechanism. In this, the content transfer is performed between structure formats of data (i.e. Table) to a set of sentences. Dataset is full focused on the Malayalam language literature authors. The set of sentences that are generated are perfectly arranged to form a biography. Here the table is a wikipedia infobox that contains various key features to represent a person. From this limited salient information, some rules are applied to effectively transfer the features into a set of statements without losing the relevant information. These statements were arranged to form a perfect biography. The proposed system's performance is analyzed through two concepts i.e. readability scores and manual human analysis. From the above concepts, an accuracy of 90% is obtained along with a college grade level standard readability score. This proves that system delivers a standard form of readable biography information without much information loss. In future work, it shall be tried on different neural network concepts. There is another possibility to extend the dataset size from Malayalam based wikipedia infobox for better coverage of information.

REFERENCES

- [1] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," arXiv preprint arXiv:1801.10198, 2018.
- [2] A. Morales, V. Premtoon, C. Avery, S. Felshin, and B. Katz, "Learning to answer questions from wikipedia infoboxes," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1930–1935, 2016.
- [3] R. Lebrecht, D. Grangier, and M. Auli, "Neural text generation from structured data with application to the biography domain," arXiv preprint arXiv:1603.07771, 2016.
- [4] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," Journal of Artificial Intelligence Research, vol. 61, pp. 65–170, 2018.
- [5] L. Sha, L. Mou, T. Liu, P. Poupart, S. Li, B. Chang, and Z. Sui, "Order-planning neural text generation from structured data," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [6] T. Liu, K. Wang, L. Sha, B. Chang, and Z. Sui, "Table-to-text generation by structure-aware seq2seq learning," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [7] P. Nema, S. Shetty, P. Jain, A. Laha, K. Sankaranarayanan, and M. M. Khapra, "Generating descriptions from structured data using a bifo-cal attention mechanism and gated orthogonalization," arXiv preprint arXiv:1804.07789, 2018.
- [8] R. Puduppully, L. Dong, and M. Lapata, "Data-to-text generation with content selection and planning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6908–6915, 2019.
- [9] A. P. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, and D. Das, "Totto: A controlled table-to-text generation dataset," arXiv preprint arXiv:2004.14373, 2020.
- [10] J. Bao, D. Tang, N. Duan, Z. Yan, M. Zhou, and T. Zhao, "Text generation from tables," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 311–320, 2018.
- [11] K. J. Miller and R. L. Rowley, "Rule based document generation system," Aug. 29 1995. US Patent 5,446,653.
- [12] S. Chen, J. Wang, X. Feng, F. Jiang, B. Qin, and C.-Y. Lin, "Enhancing neural data-to-text generation models with external background knowledge," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3013–3023, 2019.
- [13] A. M. Hein, B. Yannou, M. Jankovic, and R. Farel, "Towards an automatized generation of rule-based systems for architecting eco-industrial parks," in International Conference on Research into Design, pp. 691–699, Springer, 2017.
- [14] H. Ambavi, A. Garg, M. Sharma, N. Nitiksha, J. Choudary, R. Sharma, M. Singh, "Biogen: automated biography generation", in 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp 21-24 . IEEE 2019.
- [15] S. Chen, "A general model for neural text generation from structured data," E2E NLG Challenge System Descriptions, 2018.



**Impact Factor:
7.512**



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

📞 9940 572 462 📞 6381 907 438 ✉ ijirset@gmail.com



www.ijirset.com

Scan to save the contact details