Generating Descriptions of Structured Data using Deep Learning

Prof.Anita Shinde
Faculty of Department of Computer Engg,
Marathwada Mitra Mandal's College of Engineering
Pune, India

Kaustubh Devkar
UG Student, Department of Computer Engg
Marathwada Mitra Mandal's College of Engineering
Pune, India

Parmeshwar Matkar
UG Student, Department of Computer Engg,
Marathwada Mitra Mandal's College of Engineering
Pune, India

Sourabh More
UG Student, Department of Computer Engg
Marathwada Mitra Mandal's College of Engineering
Pune, India

Hemant Oza
UG Student, Department of Computer Engg
Marathwada Mitra Mandal's College of Engineering
Pune, India

Abstract: Natural Language generation is an emerging field of AI. It deals with generating natural language responses that can be easily understood by humans. Understanding structured data and generating reports based on this analysis is an important task for major corporations. In this paper, we try to present various techniques that are used to perform this task. We are focusing on deep learning based technique to convert structured data into its equivalent natural language descriptions. Recent advancement in deep learning led usage of models like encoder–decoder mechanism along with attention to help decoder understand where to focus exactly. We discuss the data preparations methods involved as well as various attention models used for performing this task.

***Keywords* - Artificial Intelligence, Natural Language Generation, Machine Learning, Deep Learning.**

I. INTRODUCTION

The amount of data worldwide grows fast every day. There are many problems associated with this phenomenon, be it storage, processing or proper usage. In this flood of information, it is hard to decide how to present it, pick relevant or interesting pieces, or summarize it. Often, there is a need to convert the data stored in a structured format such as tables or knowledge graphs into a form which allows easy interpretation and provides comfort to the user. The fields which deal with this problem are for example data visualization or natural language generation.

Before going into actual models, first there is need to understand our data. Structured data comes in various formats, it can be either key, value pair or a table consisting of rows and column. These columns represent the type of value and rows represent each record. Hence this problem is very different from the ones we usually see text-based deep learning. In this survey, we try to analyze what are the major techniques that are currently being utilized to fulfill this task. Firstly, looking at the dataset available for training models. Then we look at how data is prepared for feeding it to the model followed by different models that are utilized for making predictions. Finally, we see the different performance measures or evaluation techniques that are utilized to measure the performance of the model.

II. RELATED WORK

Preparing Data:

- **Categorical Embedding's**: embeddings are a great way of representing natural language data for deep learning tasks. But in our task, we have structured data where the collection of words does not represent sentences but it represents a record. These words are categorical data (e.g. for an attribute Sky –clear, not clear are two categories). Hence we can have categorical embeddings representing these categorical data. It can also be termed as the entity embeddings [10].
- **Encoding based on attribute type**: Attributes were encoded based on the attribute type. Numbers are encoded in binary representation. The Record type is encoded as a one-hot vector. Mode attribute is encoded using specific ordinal encodings [1].

Models:

Models Used for this task mostly contains the encoder– decoder architecture and attention mechanism. Encoder Decoder architecture is the standard for the task involving natural language data, but there is variety when it comes to the attention mechanism that is being utilized. Each of these attention mechanisms has its pros and cons. Different models are as follows:

- **Selective generation using LSTM and coarse fine alignment**: This is a neural encoder-aligner-decoder based model for the selective generation. It uses LSTM-based RNN for encoder and decoder. Encoder LSTM-RNN takes as input the event’s record and returns the sequence hidden state. These hidden annotations are passed to a coarse to fine aligner which is followed by a decoder. This paper uses weathergov dataset and its primary dataset and Robocop dataset to check its generalization. On weathergov dataset this model achieves an F1 score of 76:28, cBLEU score of 65:58 and sBLEU score of 75:78 [9].
- **Mixed hierarchical based attention for encoder –decoder mechanism**: This is also encoder –decoder based approach but uses a mixed hierarchical based attention mechanism. Here author used GRU-RNN. This paper contains static record attention and dynamic record attention for decoder. This paper encodes the input data based on attribute type. This paper also uses WeatherGOV dataset and achieves cBLEU score of 79.3 and sBLEU score of 87.08 [1].
- **Order planning neural text generation from structured data:** This model also presents the encoder decoder mechanism but with content based attention, link based attention and hybrid attention. This models is trained with WikiBio dataset. On WikiBio dataset this model achieves a BLEU score of 43.91 [8].
- **Structure aware seq-2-seq learning:** This model also uses the encoder – decoder mechanism but with dual attention and local and global addressing. It works on WikiBio dataset. On WikiBio dataset it achieves BLEU score of 44.89 [7].
- **Bifocal Attention Mechanism and Gated Orthogonalization:** This model works on WikiBio Dataset. It uses neural components for fused bifocal attention and gated Orthogonalization to address stay on and never look back behaviour while decoding. The maximum BLEU score achieved is 33.6 for arts category and for sports it is 52.4 [2].

Evaluation method (BLEU Score):

The Bilingual Evaluation Understudy is a score for comparing a candidate translation of text to one or more reference translations. Although developed for translation, it can be used to evaluate text generated for a suite of natural language processing tasks.

Various papers and their key findings :

Sr. No.	Title of Paper	Key Findings
1.	A Mixed Hierarchical Attention based Encoder Decoder Approach for Standard Table Summarization	Attempted to solve the problem of standard table summarization by using the hierarchical nature of tables with fixed schema. Suggested mixed hierarchical attention model with encode-attend-decode paradigms. In this approach, there is static attention on the attributes to compute the row representation followed by dynamic attention on the rows, which is subsequently fed to the decoder [1].
2.	Generating Descriptions from Structured Data Using a Bifocal Attention Mechanism and Gated Orthogonalization	Presented a model for generating natural language descriptions from structured data. Proposed neural components for fused bifocal attention and gated orthogonalization to address stay on and never look back behaviour while decoding [2].
3.	Neural Text Generation from Structured Data with Application to the Biography Domain	Shown a model that can generate fluent descriptions of arbitrary people based on structured data. Local and global conditioning used which improves model [13].
4.	A General Model for Neural Text Generation from Structured Data	Presented a general model for Data2Text. This model is built on the attention sequence-to-sequence model with three additional components: structured data embedding, copy mechanism and coverage mechanism [3].
5.	Summarizing source code using a neural attention model	Presented CODE-NN, an end-to-end neural attention model using LSTMs to generate summaries of C# and SQL code by learning from noisy online programming websites [12].

Table 4.1: Liteure Survey

III. PROPOSED SYSTEM

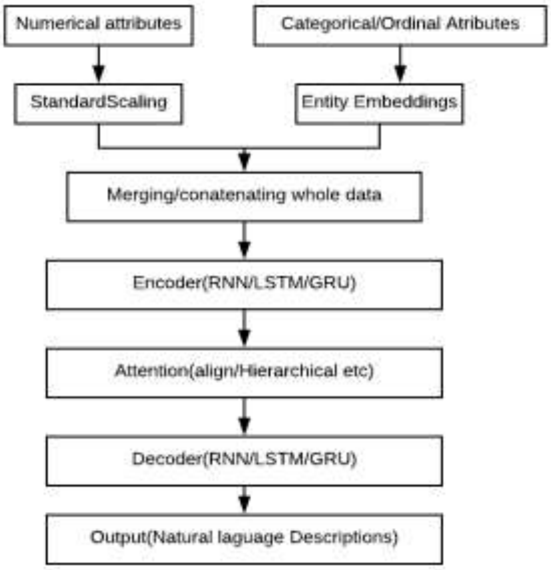
Numerical attributes
Categorical/Ordinal Atributes
StandardScaling
Entity Embeddings
Merging/conatenating whole data
Encoder(RNN/LSTM/GRU)
Attention(align/Hierarchical etc)
Decoder(RNN/LSTM/GRU)
Output(Natural laguage Descriptions)

Figure 3.1: Proposed System Architecture

In structured table data, we have columns representing different attributes. These attributes can be of type numerical and categorical. If the attribute is numerical then we need to perform standard scaling and if it is of type categorical or ordinal then we can have entity embedding representing each class values of these attributes. Then this pre-processed data gets merged into one vector and then we feed it to the encoder. After this, we can have attentions of different type depending on the type of data and performance required. Finally, we start decoding the sequence, which gives us the natural language description of the data.

Available Datasets:

There are two major datasets available for training the model for this task. These are:

- **WeatherGOV dataset:** The weather forecasting dataset consists of 29528 scenarios, each with 36 weather records (e.g. temperature, sky cover, etc.) paired with a natural language forecast (28:7 avg. word length).

- **ROBOCUP dataset:** This dataset consists of only 1539 pairs of temporally ordered robot soccer events (e.g., pass, score) and commentary drawn from the four-game 2001–2004 RoboCup. Each scenario contains an average of 2.4 event records and a 5.7 word natural language commentary.

- **WIKIBIO Dataset:** WIKIBIO dataset which contains 728,321 biographies from WikiProject Biography3 (originally from English Wikipedia, September 2015.Each data sample comprises an infobox table of field content pairs, being the input of our system. The output sentence typically serves as a summary of the article. In fact, the target sentence has 26.1 tokens on average, which is actually long. Also, the sentence contains information spanning multiple fields.

IV. CONCLUSION

Most of the models that we studied are based on Encoder Decoder approach. While preparing the data for model building Categorical Embedding’s and Encoding based on attribute type are used. Attention mechanism is used in the model for getting the output of dense layer using Softmax function.LSTM-based RNN ,GRU- RNN are used for encoder and decoder. All the models are evaluated based on the BLEU score. The proposed system can be used to generate textual descriptions of structured data.

V. FUTURE WORK

As future work, it is possible to tackle general tabular summarization where the schema can vary across tables in the whole dataset. The system can be improved for unstructured data.

REFERENCES

[1] Parag Jain, Anirban Laha, Karthik Sankaranarayanan, Preksha Nema, Mitesh Khapra, Shreyas Shetty, "A Mixed Hierarchical Attention based Encoder Decoder Approach for Standard Table Summarization", Proceedings of NAACL-HLT 2018, Association for Computational Linguistics,(June-2018),pages 622–627.

[2] Preksha Nema, Parag Jain, Anirban Laha, Karthik Sankaranarayanan, Mitesh Khapra, "Generating Descriptions from Structured Data Using a Bifocal Attention Mechanism and Gated Orthogonalization", Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, June 1 - 6, 2018, pages 1539–1550.

[3] Shuang Chen, "A General Model for Neural Text Generation from Structured Data", Association for the Advancement of Artificial Intelligence (www.aaai.org). (2018).

[4] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. 2017, "Image-to-markup generation with coarse-to-fine attention", In Proceedings of the 34th International Conference on Machine Learning,ICML2017,Sydney,NSW,Australia,6-11 August 2017, pages 980–989.

[5] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. "Challenges in data-to-document generation". In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2253–2263.

[6] Jeffrey Lingand, Alexander M. Rush. 2017, "Coarse-to fine attention models for document summarization", In Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017, pages 33–42.

[7] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang and Zhifang Sui, "Table-to-text Generation by Structure-aware Seq2seq Learning",Nov 27,2017, arXiv

[8] Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, Zhifang Sui, "Order-Planning Neural Text Generation From Structured Data", Sept 1,2017, arXiv

[9] Hongyuan Mel, Mohit Bansal, Mathew R. Walter, "What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment", Jan 8,2016, arXiv.

[10] Cheng Guo and Felix Berkhahn.2016."Entity Embeddings of Categorical Variables" , CoRR, April 25,2016, arXiv.

[11] Emilie Colin, Claire Gardent, Yassine Mrabet, Shashi Narayan, and Laura Perez-Beltrachini, 2016, "The webnlg challenge: Generating text from dbpedia data", In INLG2016-ProceedingsoftheNinthInternational Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK, pages 163– 167.

[12] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer, 2016, "Summarizing source code using a neural attention model", In Proceedings of the 54th Annua lMeeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2073–2083, Berlin, Germany. Association for Computational Linguistics.

[13] Remi Lebret, David Grangier, Michael Auli, "Neural Text Generation from Structured Data with Application to the Biography Domain", the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, November 1-5, 2016, pages 1203–1213.

[14] Ioannis Konstas and Mirella Lapata. 2013 "Inducing document plans for concept-to-text generation" In EMNLP. ACL, pages 1503–1514.

[15] Gabor Angeli, Percy Liang, and Dan Klein. 2010, "A simple domain-independent probabilistic approach to generation", In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing .Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '10, pages 502–512

[16] Albert Gatt and Anja Belz. 2010 "Introducing shared tasks to nlg: The tuna shared task evaluation challenges" , pages 264–293