GENERATION OF TEXT FROM STRUCTURED DATA USING GENERATIVE MODELS

Anusha Prakash
anushap@andrew.cmu.edu

Mansi Gupta
mansig1@andrew.cmu.edu

Raghuveer Chanda
rchanda@andrew.cmu.edu

Srividya Potharaju
spothara@andrew.cmu.edu

ABSTRACT

Neural models have shown significant progress in the generation of short descriptions of text over unstructured data. We wish to tackle text generation using structured data like tables, infoboxes, databases, etc to present meaningful information along with the capability to control the text semantics like tense and length. We will use the WIKIBIO dataset which contains around 70k articles from Wikipedia for the same. We will use an encoder-decoder model with dual attention and copy mechanism paired with a generative model that is used to control the semantics.

1 INTRODUCTION

A classic problem in Natural Language Generation (NLG) is generating text that adequately and clearly describes the structured data, such as table, graph or infobox or any other knowledge base. According to Mei et al. this form of NLG imposes two goals, *what to say*, the selection of relevant subset of the input data to present, and, *how to say it*, the surface realization of a generation. To realize these goals, content selection (CS), content ordering (CO) and precisely generating the relations (RG) from the database are of utmost importance. These three aspectsre the basic requirements of such an NLG system, and hence, form criterion for evaluation of text generation using structured input(Wiseman et al.) In this work, in addition to satisfying these objectives, we wish to impose explicit semantic constraints over the generated text, like controlling the length, tense or voice of the output.

In this work, we consider the problem of generating coherent biographical description of people using Wikipedia infoboxes as input. Infoboxes are in the form of collection of fields (Name, Born, Nationality etc) and corresponding value(s). Given the infobox of a living person, say Barack Obama, the generated biography should be in present tense as opposed to generating biography of say, Mahatma Gandhi, which should be in past tense. Similarly, one might want to keep the generated biography short and extract only the most important sentences. This information can either be inferred from the infobox or explicitly provided to the model. We plan to exercise this control by using generative models like Variational Auto Encoders (VAEs) or Generative Adversarial Networks (GANs) in addition to commonly used sequential encoder-decoder modules.

2 BASELINE OR INITIAL EXPERIMENTS

For the baseline, we plan to implement standard encoder-decoder based approach incorporating dual attention (global and local attention) and copy mechanism Wiseman et al.. We plan to experiment

with joint copy, conditional copy and reconstruction loss based approaches mentioned in this work.

Another obvious baseline to implement is structure aware seq2seq based approach as mentioned in Liu et al.. In addition to dual attention in the decoder, they also encode field information by adding a field-gate to the cell state of the LSTM unit.

3 Proposed Approach

The base model that would take structured data as input and output unstructured data, would be an encoder-decoder model where the decoder will have dual local and global attention. The encoder is used to embed the input relations, for which an MLP would be sufficient. The decoder would be a sequential RNN based model, with LSTM or GRU cells where we also incorporate a copy mechanism as proposed by Gu et al. to cope with rare words.

In addition to encoder-decoder modules, a generative module will be used to control the characteristics of the output. We plan to use the technique of disentangling latent space representations to incorporate the characteristics we desire using a modified version of VAE. This approach is heavily inspired by the work of Hu et al.. The discriminator of the generative module will control the required characteristic and train the generator. We will use reconstruction loss to backpropagate through the entire pipeline. Please note that the encoder-decoder and the generator modules are not disconnected, rather the loss will backpropagate to the input.

For evaluation, we will use the extractive evaluation techniques mentioned in Wiseman et al. to compare Content Selection (CS), Content Ordering (CO) and Relation Generation (RG) between generated and expected description. We also want to explore the possibilty of employing adversarial evaluation approaches. For evaluating the output controlled by length, we will compare it against only the first few sentences of the true output. The tenses would already be captured correctly in the biography, and hence, controlling the tenses should ideally decrease the loss organically.

4 Related Work

To encode both the content and the structure of a table, Liu et al. propose a novel structure-aware seq2seq architecture which consists of field-gating encoder and description generator with dual attention. They experiment on the WIKIBIO dataset and show that their model is capable of generating coherent and informative descriptions based on the comprehensive understanding of both the content and the structure of a table.

Chisholm et al. investigate on generating single sentence descriptions from facts derived from Wikidata slot-value pairs. They train a RNN sequence-to-sequence model with attention to select facts and generate textual summaries and obtain a performance much better than the vanilla sequence-to-sequence model.

Lebret et al. propose a model which builds on the conditional neural language models, by employing copy actions to deal with the large vocabulary. To account for the structured data, the model is allowed to embed words differently depending on the data fields in which they occur and obtain a model that outperforms a Templated Kneser-Ney language model by nearly 15 BLEU.

Liang et al. present a generative model that simultaneously segments the text into utterances and maps each utterance to a meaning representation grounded in the world state. They propose a probabilistic generative model, hierarchical hidden semi-Markov model that treats text segmentation, fact identification, and alignment in a single unified framework and show that their model can generalize across three domains of increasing difficultyRobocup sportscasting, weather forecasts (a new domain), and NFL recaps.

Previous work in the domain of NLG includes template-based, neural model based, copy and reconstruction based approaches. Although neural systems are quite standard at generating clear outputs, they perform quite poorly in terms of content selection and capturing long-term structure

on a more challenging dataset like ROTOWIRE. Infact a template-based approach performs much better according to the evaluation by Wiseman et al.. Though using copy-based models and reconstruction terms along with neural models boosts the BLEU performance, the quality of the generated text is still far from human-level output. This is the motivation for our work to experiment with a novel approach of employing generative models for this purpose.

5 Data & Technical Requirements

We intend to employ the WIKIBIO dataset proposed in Lebret et al. as the benchmark dataset. WIKBIO contains 728,321 articles with corresponding infoboxes from English Wikipedia. There is also some previous work on some smaller datasets which contain few tens of thousands of records like WEATHERGOV by Liang et al. and ROBOCUP by Chen & Mooney. WIKIBIO has more amount of data and involves generating text by understanding the structure and content of the data.

A more challenging dataset we also want to focus on are ROTOWIRE and SBNATION by Wiseman et al.. The first dataset, ROTOWIRE, consists of medium length game summaries written by professionals and are well structured with game statistics. The second dataset, SBNATION, consists of significantly larger length game summaries written by fans but are more challenging as the language is not consistent across.

References

David L Chen and Raymond J Mooney. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pp. 128–135. ACM, 2008.

Andrew Chisholm, Will Radford, and Ben Hachey. Learning to generate one-sentence biographies from wikidata. *arXiv preprint arXiv:1702.06235*, 2017.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pp. 1587–1596, 2017.

Chloé Kiddon, Luke Zettlemoyer, and Yejin 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.

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*, 2016.

Percy Liang, Michael I Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 91–99. Association for Computational Linguistics, 2009.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. *arXiv preprint arXiv:1711.09724*, 2017.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. *arXiv preprint arXiv:1509.00838*, 2015.

Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. Order-planning neural text generation from structured data. *arXiv preprint arXiv:1709.00155*, 2017.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.