Ontological Enrichment of Pre-Trained Language Models for Data-to-Text Generation

Ashish Bharadwaj Srinivasa

October 2020

CS224U Natural Language Understanding

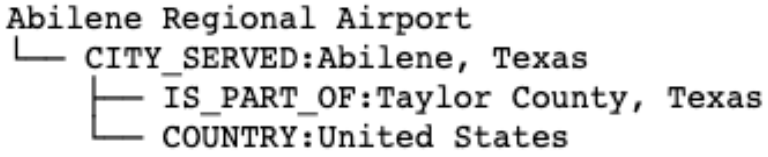
Experiment Protocol

Hypotheses

Data-to-Text generation has gotten a lot of interest in recent years due to its direct applicability to question answering systems, chatbots, summarization, etc. Moreover, most enterprise data is highly structured and stored in the form of databases and knowledge bases. Most approaches for table-to-text conversion work with flat ontological structures by linearizing the table cells and using state-of-the-art language models for text generation. In this work, we aim to prove that incorporating rich hierarchical ontologies present in tables as semantic triples generates more semantically sound text given a set of highlighted table cells. We hypothesize that pre-trained language models combined with ontology encodings will capture the rich semantic interdependencies among values in the tables. Lastly, we hope to build a generalized model that effectively handles out of domain semantic triples.

Data

For proving the above hypotheses, we plan to use the DART [1] dataset curated by Radev et al. DART is an open-domain data-to-text dataset gathered from highlighted table cells. The data is in the form of semantic RDF triples and their corresponding verbalized representation. What makes DART unique and conducive to the above hypotheses is its emphasis on capturing hierarchical ontological structures. While previous table-to-text corpora such as ToTTo [2], WikiTableText [3], etc. capture only the flat ontological structures of table rows and columns headers, DART captures rich semantic interdependencies among these table cells as RDF triples. The below figure shows an example hierarchical dependency among table cells shown in DART.



Additionally, DART also places strong emphasis on its open-domain nature. While popular datasets like WebNLG [4], E2E [5] contain data from limited number of semantic domains, DART captures tables extracted from Wikipedia tables in general. This poses an added challenge and forces the models to be generalized enough to handle a wide variety of domains.

The data is pre-split into train, dev and test sets with each data point containing a list of semantic triples in JSON/XML format carefully collected by a combination of human and automatic annotation. There is a total of 82,191 examples generated from a total of 5,623 tables. The two main challenges posed by DART as described above are due its open-domain nature and hierarchical ontology structure. This leads to popular state-of-the-art table-to-text methods being less effective as they only consider a linearized (or flattened) semantic representation. Thus DART is the ideal dataset to prove our hypotheses.

Finally, to truly measure the out-of-domain generalization of the model trained, we plan to test the models on the WebNLG dataset, which has a held out test set containing data from completely unseen domains. DART training has proven to be effective for out-of-domain generalization and we hope to measure this impact when coupled with rich ontological representations. Both DART and WebNLG are publicly available and ready for download.

Metrics

For the data-to-text generation task described above, we use two types of metrics - sequence correlation based metrics and semantic match/distance metrics. Improving both these two types of metrics is critical for the task we are solving and the hypotheses proposed. The quality of any text generation task will first need to be evaluated on a purely token matching basis to measure how correlated the generated text sequence is to the true/expected sequence. These can be captured by using the popular sequence matching metrics such as BLEU [6] and METEOR [7]. Both the metrics are a modified version of the precision metric that allows comparing the predicted text with the reference text on a token co-occurrence level. While, BLEU is a popularly used metric that most state-of-the-art models report on, we additionally use METEOR as it tackles some of the problems with BLEU. Specifically, METEOR considers matching on word stems as opposed to exact token match and also allows for comparing sentences directly as opposed to corpora.

Secondly, we want to measure how semantically close the generated text is to the reference text. This is even more important for our specific hypotheses as we want to measure if the rich semantic representation from the hierarchical ontologies translated to the generated text. Pure token matching approaches might miss the semantic distinction from hierarchical information, due to their inherent unordered nature. To track this we use BERTScore [8] and BLEURT [9]. These are machine learned metrics which compare a candidate sentence to a reference sentence by computing similarity at a token level using contextual embeddings as opposed to exact matches like the previous two metrics. This allows the metrics to capture rich semantic similarity assessments and makes them very effective for evaluating our hypothesis. Additionally, as these metrics are fully differentiable there is scope to use these metrics as part of the loss function to optimize the model as well.

Finally, we also want to evaluate the generated text qualitatively as well. This allows us to quickly track how fluent (semantically sound) a sentence is, how faithful the sentence is to the source semantic triples and the coverage of the triples in the text. This analysis will allow us to perform error analysis and helps us design better metrics to capture this information quantitatively. There is scope to introduce new machine learned metrics as well that capture coverage of hierarchical information in the generated text as we do the qualitative analysis.

Models

Radev et al. have established the baseline metrics on DART for data-to-text generation on these 3 models:

- Seq-to-Seq with Attention → Simple RNN based encoder decoder architecture with attention mechanism.
- Transformer [10] → end-to-end transformer based model that generates text from linearized triples
- BART [11] → pre-trained sequence-to-sequence language model using a transformer based architecture which has a BERT encoder and a GPT like left-to-right decoder

	BLEU ↑	METEOR ↑	TER ↓	MoverScore ↑	BERTScore(F1) ↑	BLEURT ↑
End-to-End Transformer	19.87	0.26	0.65	0.28	0.87	-0.20
Seq-to-Seq with Attention	29.60	0.28	0.62	0.32	0.90	-0.11
BART	37.06	0.36	0.57	0.44	0.92	0.22

Kale et al., [12] currently has the state-of-the-art model for data-to-text generation on most other datasets by fine tuning a pre-trained language model called T5, which was originally trained for text-to-text generation. We plan to fine tune and evaluate the T5 model on DART. Lastly, we also hope to fine tune the BERT-to-BERT [13] model on DART, this model combines a BERT encoder and decoder. This model has proven to be very effective on the ToTTo dataset with appropriate linearization. Overall, we will fine tune the pre-trained language models, namely BART, T5, BERT-to-BERT, which have proven effective on other datasets and text generation tasks.

To enhance the above models, we hope to augment the embeddings with the existing hierarchical ontological structure in the semantic triples. This can be approached in two ways - effective linearization strategies and architecture level modifications. Firstly, the current models do not capture the hierarchical ontology because the RDF triples are linearized before being fed into the model as a sequence. We should address this. But since we are pursuing an end-to-end approach we do not want to learn a sentence ordering model here but instead simply capture the hierarchical structure of the RDF triples. Secondly, we can make architectural changes in the encoder and decoder components by extending the idea of hierarchical attention [14] from words-sentence to RDF triples→ontology. The interesting challenge to tackle here will also be the unordered nature of the RDF triples itself, which we hope to do by using the trick proposed in Set Transformers [15] where the transformer model is tweaked to work for unordered input by removing the positional encodings. The idea is to capture the positional information within an RDF triple and ignore the overall ordering. This also allows us to augment the linearized input data by randomly shuffling the order to help fine tuning. Lastly, we can learn a fidelity classifier that explicitly scores the generated text based on its correlation with the input hierarchical ontology. This technique will penalize predictions which internally flatten the hierarchy or ignore interdependencies and thereby improve the conditioning of the generated text on the ontological structure.

General Reasoning

The core idea behind the hypothesis is that data-to-text generation models can produce much more semantically sound sentences if they were given additional ontological information. As a secondary goal, we also hope to make the models as generalized as possible in order to quickly fine tune to new domains. For these reasons, we have selected the DART dataset which is both open-domain and contain hierarchical semantic triples. To tackle the generation, we have primarily chosen to start with pre-trained language models because these models have been trained on really huge amounts of text data. This makes them especially effective when faced with out-of-domain semantic triples as they can still generate semantically sound output sentences. Secondly, to address the handling of hierarchical input information, we propose revising the linearization procedure currently used as it explicitly flattens the ontology before feeding into the language model. Alternatively, we also propose to make architectural changes in the encoder and decoder blocks to enhance the pre-trained language models to incorporate the hierarchical structure of the inputs. Finally, we want to address two main concerns with current approaches - the forced ordering produced by current linearization methods and the lack of explicit penalization for when the model fails to incorporate the hierarchical information. With these improvements, we believe the models produced will be better than the baseline models established on DART by being more semantically sound as measured by the BERTScore and BLEURT metrics and will tightly incorporate the ontology structure in the generation.

Summary of Progress

Until now, we have strictly defined the task we are attempting to tackle and reviewed the literature for the most popular and effective ways to solve it. After analyzing the moving trend towards end-to-end systems, we have decided to pursue an end-to-end approach and formulated the core hypothesis of the work → richer ontological representation should improve data-to-text generation. We have identified and downloaded the DART dataset for this research. We have performed some

exploratory data analysis to get a feel for the data, the hierarchical nature of the triples, the textual outputs, distribution of triples, etc. We have also worked on writing a parser to convert the data into a suitable training form using with BPE. We have tested the BERT, BART and T5 models from huggingface transformers library on dummy data and are working on linearizing the inputs to replicate the baseline results and evaluate the new models. We have also written pipelines to evaluate generated text to produce BLEU, METEOR, BERTScore, BLEURT metrics. Going forward, we plan to first set up pipelines to quickly tweak and evaluate the pre-trained models mentioned. Then we plan to tweak the different linearization strategies and model architectures proposed in the models section above and systematically analyze the metrics. Additionally, we hope to perform qualitative evaluation to find problems with the current models and devise quantitative metrics from the findings if needed. We will also qualitatively try to determine triples for which ontological enrichment works well versus where it degrade the generated text. One of the challenges we anticipate are lack of time for hyper parameter optimization to tweak the models.

References

[1] Dragomir Radev et al., 2020, “DART: Open-Domain Structured Data Record to Text Generation”
[2] Ankur P. Parikh et al., 2020, “ToTTo: A Controlled Table-To-Text Generation Dataset”
[3] Junwei Bao et al., 2018, “Table-to-Text: Describing Table Region with Natural Language”
[4] WebNLG Challenge, https://webnlg-challenge.loria.fr/challenge_2017/
[5] E2E NLG Challenge, http://www.macs.hw.ac.uk/InteractionLab/E2E/
[6] Kishore Papineni et al., 2002, “BLEU: a Method for Automatic Evaluation of Machine Translation”
[7] Banerjee, S. et al., 2005, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments"
[8] Tianyi Zhang et al., 2020, “BERTScore: Evaluating Text Generation with BERT”
[9] Thibaut Sellam et al., 2020, “BLEURT: Learning Robust Metrics for Text Generation”
[10] Ashish Vaswani et al., 2017, “Attention Is All You Need”
[11] Mike Lewis et al., 2019, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”
[12] Mihir Kale, 2020, “Text-to-Text Pre-Training for Data-to-Text Tasks”
[13] Sascha Rothe et al., 2020, “Leveraging Pre-trained Checkpoints for Sequence Generation Tasks“
[14] Zichao Yang et al., 2018, “Hierarchical Attention Networks for Document Classification”