Sequencing Legal DNA: Project Outline

Partisan Responses

July 14, 2020

1 Motivation

This project aims to produce a question answering system capable of generating answers in the style of republicans and democrats, similar to those from the U.S. Congressional Record. This is inspired by recent research in natural language processing focused on producing ”controlled” text generators. Rather than merely producing coherent text, the current research aims to develop means of generating text that has some particular attributes such as politeness or positiveness. The attribute we want to study is political slant. Several studies have shown that the particular wording of a text can reveal the author’s political leaning. For instance, what democrats refer to as ”estate tax” is called ”death tax” by the republicans. We hypothesise that, when asked the same question, a republican would answer differently from a democrat. Therefore, we aim to develop a system that produces answers in the style of either republican or democrats, while preserving factual information.

2 Literature Review

2.1 Knowledge Graph Augmentation

The paper from Chen et al. (2017) gives insights about how to construct a political opinion-aware knowledge graph based on an existing knowledge base by incorporating ideological attributes on entities in the knowledge base. It uses a lexicon-based approach to estimate ideological scores of entities in input text and propagates to other entities in the knowledge base according to defined criteria.

2.2 Models in General

From Gui et al. (2020) we can get good insights on how to train GAN models from the ground up and for our project more important, the possible problems a GAN faces, such as mode collapse, where always the same text is generated, since it is considered best. Also a a possible measurement, not relying on NLP, would be inception score.

From the scoreboard of typical NLP tasks in Ruder (2020) we can get datasets which could be used to fine-tune a model. For our project, the most interesting benchmarks are the text generation aspect of shARC, the type of questions that QuAC answers and the format of the ground data of QALD.

2.3 Text Generation

A desirable characteristic of a question answering system is the ability to generate answers that are consistent with the question, without drifting to another topic. Yang et al. (2019) show that integrating facts from a knowledge base is useful in topic-to-essay generation as it helps generate texts that are more novel, diverse and topic-consistent. They propose a memory-augmented neural model trained using adversarial learning that incorporates knowledge from ConceptNet. The generator is an encoder-decoder model that takes a list of topics as input and produces text consistent with the topics by using the attention mechanism to find the most relevant concepts from the memory matrix. The discriminator is used to evaluate whether the generated text is consistent with the input topics. Similarly, Koncel-Kedziorski et al. (2019) applies a state-of-the-art information extraction system that is used to construct knowledge graphs on texts with implicitly strong structural format. In the study, such graph and text features are encoded via Graph Transformer and Bi-RNN in training stage. Attention-based decoder then generates an abstract of scientific papers from encoded embeddings.

The objective of the paper Li et al. (2017) is to create a GAN, where the discriminator performs a Turing test. The generated texts should be as human-like as possible and they define it in three qualities: informativity, coherence and ease of answering. Using these qualities, this project cares most about informativity and coherence. As model they started with a seq2seq model and improved on that by using tf-idf vectors and keeping track of previously used words, to have a lower chance of contradicting itself. They also had the interesting problem, that one model scored very good by outputting incoherent random words and they showed that by establishing a measure that scores on the capability to distinguish between generated text and random text.

The method presented in Clark et al. (2018) heavily leans on named entities to create (children's) stories. Unfortunately, we can not replicate the same method, since for this project we are much more bound to knowledge. A potential use of this paper would be to instead of a new named entity, a new node of a knowledge graph gets introduced.

In terms of hierarchical text generation, the paper Fan et al. (2018) partly aims to resolve long-range dependency and creativity in story generation by designing a convolutional Seq2Seq model with self-attentive decoder based on a prompt trained from convolutional language model. The generation process uses top-k random sampling for words to avoid repetitiveness (in beam search) and unlikely words (in completely random sampling).

Due to the unsupervised nature of text generation, the evaluation stage tends to involve much effort from human experts. Therefore, automatic evaluation metrics is a demanding foundation to achieve reasonable and promising results. Diversity and creativity in dialogue response imposes challenges in current unsupervised evaluation metrics. The survey from Liu et al. (2016) examines pros and cons of two mainstream methods-word overlap-based(BLEU, ROUGE, METEOR) and word embedding-based (Greedy Matching, Embedding Average and Vector Extrema)-in such text generation tasks. The empirical study shows a low correlation between these metrics and human judgements in certain tasks, and warns against overly usage of them without thorough consideration into specific task. Without better recommendations, it encourages variations in word embedding-based method to evaluate dialogue responses. In the case of graph-to-text task, Sheffer et al. (2019) tries to leverage a novel metric by comparing input knowledge graph with an extracted graph from generated text. In evaluation procedure, this method reduces the dependency of reference texts which are common practices in other metrics. From the experiments, this method promisingly has a higher correlation with human evaluation than existing popular metrics like BLEU.

2.4 Style Transfer

Style transfer in language has received a lot of attention recently. The key challenge is preserving the semantic content of the generated text while controlling aspects such as sentiment, tense, gender or political slant. In order to achieve this, Xu et al. (2018) use a cycled reinforcement learning technique that does not require paired samples for performing sentiment-to-sentiment translation. Their system employs a neutralisation module that extracts the non-emotional words from the input and an emotionalisation module that reconstructs the original sentence. They use policy gradient reinforcement learning for training the model and use rewards for content preservation and sentiment confidence. Prabhumoye et al. (2018) perform style transfer through back translation and evaluate their technique on 3 style transformations, including political slant. They use a neural machine translation translation encoder to transform the original sentences from a source language to a target language, hypothesising that translation helps eliminate the stylistic content while preserving the semantic content. In addition, they train separate decoders for each style, guided by style classifiers, to back-translate to the source language. Hu et al. (2017) develop an enhanced VAE with an extended wake-sleep method for producing text with a certain attribute such as sentiment or tense. Their method uses a generator for producing plausible text and a discriminator that determines how well the generated text captures the desired attribute and drives the generator to produce text that is coherent with this characteristic. Peng et al. (2020) fine-tune the GPT-2 language model using a policy gradient reinforcement learning technique to generate text with a certain desirable attribute such as normativity or positivity. They use an existing normative text classifier to determine whether some generated text is normative and treat this output as a reward signal. This reward is then augmented to the cross-entropy loss used when training the GPT-2 model.

3 Dataset

The main corpus is the congressional records of the United States archived in the Stanford Social Data Science Collection. It contains daily (11.35 GB) and bound editions (32 GB) of full speeches, metadata of the speech and speaker, parsed bi-grams and more. The edition ranges from the 43rd to 111th Congresses, the bound edition from 1873 to 2011 and the daily edition from 1981 to 2017. The full speeches are maintained in files per congressional session, and within each file the speeches are parsed on day-level. The dataset also summarises partisan phrases, vocabulary and 22 debate topics in separate files.

The speech manuscripts consist of sufficient opinions and debates on diverse but familiar topics from different ideological perspectives. Such semantic and stylistic divergences of conservative and democratic speakers can guide machines to learn and generate texts toward a targeted question in designated topics. The dataset will potentially be split by speech-level, such that one speech represents one data point. The representation will incorporate ideological variables from speaker metadata, speech topic, (and partisan phrases). The training model is likely to utilise organised vocabulary from the dataset too.

The preprocessing is planned to be proceeded as below with more refinement and adjustment. First, considering the machine capability, we may sample a subset from the entire collection for training. To better augment the knowledge graph, we now believe that the corpus needs to be stemmed and lemmatized for better entity and relationship extraction. In addition, stop-words and most-irrelevant ideological words (requiring detailed definition) should be removed. Lastly, the base knowledge graph might be acquired from DBpedia or similar sources, which are widely used in relevant research papers.

4 Methods

Keep in mind that because of the size of the data and the models planned to use, it will not be possible to run the project on our laptops, but a part of the challenge will be to find and use clusters.

The task of replying with a speech to question can be divided into three sub tasks.

4.1 Knowledge Graph Augmentation

Speeches should be based on the political context, that is why it is necessary to acquire a knowledge base, which in our case will be knowledge graph, that will be enhanced with slant of different words, like in Chen et al. (2017). To successfully do so, we need to preprocess the data and have speeches connected to parties.

4.2 Answering Graphs to Questions

As we do want to answer questions with speeches, we need to find the corresponding subgraph. This can either be done by incorporating complex systems such as BERT by Devlin et al. (2018) or simpler techniques such as cosine similarity of tf-idf vectors.

4.3 Creating Speeches from Knowledge Graphs

Once we have the correct subgraph, the last sub task is to create speeches based on this graph, like it is done in Yang et al. (2019). Another challenge of this task is to find speeches, which answer the question and not just start talking after the question.

4.4 Evaluation

Good results should be (non-)partisan and also close to being human-like. Measuring these properties can be done by rather simple classification tasks and also more complex measures such as BLEU.

5 List of Tables and Figures

- Corpus statistics (e.g. number of questions, number of speeches of republicans/democrats)
- Diagram of system components
- Example for every sub step
  - Enhanced graph from a text
  - Returned subgraph from a question
  - Generated text from a subgraph
- Generated sample of answers in the style of republican, democrat and neutral for the same question
- Real sample of answers in the style of republicans for democrat questions and vice versa
- Slant of the answers compared to the slant of the question
- Table of evaluation metrics (e.g. topic consistency as done in Yang et al. (2019), BLEU score)

6 Timeline

- **By 1.06.2020** : Get familiar, with the dataset, extract speeches for republicans/democrats, build the knowledge graph
- **By 15.06.2020** : Map a question to a knowledge subgraph
- **By 10.07.2020** : Text generation based on knowledge graph
- **By 17.07.2020** : Evaluation and analysis
- **By 24.07.2020** : Finish writing up the report

7 Work division

Details on work division among group members are written in a separate document.

References

Chen, W., Zhang, X., Wang, T., Yang, B., and Li, Y. (2017). Opinion-aware knowledge graph for political ideology detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3647–3653.

Clark, E., Ji, Y., and Smith, N. A. (2018). Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2020). A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2017). Toward controlled generation of text.

Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., and Hajishirzi, H. (2019). Text generation from knowledge graphs with graph transformers.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., and Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.

Peng, X., Li, S., Frazier, S., and Riedl, M. (2020). Fine-tuning a transformer-based language model to avoid generating non-normative text.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation.

Ruder, S. (2020). Nlp-progress. [Online; accessed 4-May-2020].

Sheffer, O., Castel, O., and Landau, R. (2019). Going grean: A novel framework and evaluation metric for the graph-to-text generation task.

Xu, J., Sun, X., Zeng, Q., Ren, X., Zhang, X., Wang, H., and Li, W. (2018). Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach.

Yang, P., Li, L., Luo, F., Liu, T., and Sun, X. (2019). Enhancing topic-to-essay generation with external commonsense knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2002–2012, Florence, Italy. Association for Computational Linguistics.