Applying Transformers and NLP Computational Techniques to America in One Room

Stanford CS224N Custom Project

Haroun Ahmed
Department of Computer Science
Stanford University
haroun@stanford.edu

March 20, 2021

Abstract

This project seeks to apply a scalable, NLP based computational mechanism to automate portions of the work done by the Center for Deliberative Democracy(CDD) in their Deliberative Polling project. [3] The machine based classifier that this project is attempting to profile is intended to emulate the currently manual task of tagging transcripts of worldwide deliberations about controversial topics with the intention of 1. Extracting argument dialogue segments and 2. Assessing argument deliberative quality as per the conventions set forth by the CDD [4]. Top level, these techniques involve humans tagging transcripts using a 0, 1, 2, 2+ scale, reflecting the respective number of provided reasons to support an argument. The computational techniques that will be profiled for their capacity to replicate human tagging that this project will explore will be transformer based language models such as XLNet [5], a custom bag of words (BOW) + random-forest based classifier, and random chance. Empirical results have displayed an increase in model generalizability when transferring to the XLNet model, indicating a better capacity to emulate human transcript tagging across all categorizations, conversational subjects, and discussion locales.

1 Key Information to include

- Mentor: Andrew Wang
- External Collaborators (if you have any):
- Sharing project: No

2 Introduction

If natural language processing (NLP) is meant to utilize computational techniques to emulate natural human dialogue, then it stands to reason that there should be paradigms for extracting and constructing a method for validating an assessment for perhaps what humans do most with their language - argue. The learning and discussion based discourse concerning current and historical social and political issues that has been one of the primary topics of human communication have largely moved to online forums, or otherwise have been transcribed in text format. As such, the resultant dialogue corpora are massive, cover any topic imaginable, and provide rich data to address questions about the nature of human discussion. Naturally, the extension comes to two necessary tasks to ascertain meaningful information about the argument topic and broader human argumentation in general:

1. Argument Extraction: How can we extract segments in dialogue that raise an argument?

2. Argument Assessment: How we might assess the overall quality of a posited argument?

The goal of this project is to develop a model to automatically discover semantic aspects of arguments across multiple dialogues on a topic and predict the quality of extracted arguments, with a secondary goal of generalizing the model to be topic independent. Developing an automatic process to classify argument salience at or near human level of assessment would be a boon to the larger pursuit of the CDD's mission to research about democracy, public opinion, and the interrelationships between the two.

Moreover, we want to establish a baseline for this difficult classification problem. Even as NLP gets better at adequately representing and translating human meaning and intention, classifying argument quality seems to be a more abstract, even slightly subjective task. The CDD has developed experimental consensus within the academic community for their method, and their technique boasts quite high inter-rater reliability. Meanwhile, NLP researchers have attempted to broad strokes apply techniques towards arguments in a manner similar to comprehension and discourse. This technique that doesn't have the academic communication community backing in the same manner. This project serves as a marriage between computational and framing techniques seen in NLP with Deliberative Polling and Communication studies in general.

3 Related Work

Perhaps unsurprisingly, given the myriad of directly applicable use cases, there have been several attempts to apply NLP techniques towards argument assessment. One of the seminal papers by Swanson et al was the first to attempt to utilize computational models to raise a two-step method ahead of traditional supervised learning on labeled topic-specific data to offer domain independent argument context [6]. Specifically, the researchers extended previous linguistic research on argumentation to develop custom features to train a regressor to predict the quality of extracted arguments with RRSE (root relative squared error) of 0.72 with minimal cross domain loss implying relatively domain independence.

Swanson et al raised a litany of hand-curated features that they claim as indicators of argument presence and quality: Sentence Length(SLEN), Word Length(WLEN), Speciteller(SPTL), Kullback-Leibler Divergence(KLDiv), Discourse (DIS), Part-Of-Speech N-Grams (PNG), Syntactic and (SYN). These features were cobbled together into 3 meta-feature sets for domain specific vs cross domain testing.

The related work certainly explores an incredibly interesting and widely applicable area of research into text based argumentation. There are however portions that leave a lot to be desired - namely the classification of argument quality that authors tend to use. General proxy of argument quality is meant to reflect "how easily the speaker's argument can be understood from the sentence without any context," usually stylized as a a 0.0 to 1.0 sliding scale with an interrater reliability metric for statistical strengthening. While a statement like "gun control is good because I had a bad experience with guns in the past," would certainly be easily understood outside of context, it doesn't seem like a particularly good argument. Some followup work has been done to combine such techniques with distributional semantics, but some of the more modern NLP transformer based paradigms haven't really been utilized in combination with these empirical assessment techniques. [1]. This project is an attempt to extend some of these more modern NLP frameworks and techniques to this problem.

4 Approach

Deliberative Polling™ is a technique to illustrate how a better informed populace might affect public discourse on controversial issues. The process involves collecting a representative sample, conducting English language synchronous workshops across the world with bipartisan experts and moderated debate, and then analyzing the quality of argument by proxy of offered justifications [4]. Unfortunately, the data as collected has been optimized for ease of transcript collection and tagging for a team of human applicants rather than pre-engineered for input into an ML model, so some data engineering is required.

The baseline human coded transcript tags for each workshop are in separated .xlsx files for the respective moderated group. Custom code was engineered to take each sheet from each file, turn them into csv's, extract each of the example utterances while filtering out any tagged moderator content to create a by workshop set of examples to be used for either the BOW + Classifier or the hopefully more powerful Transformer model. The most natural baseline of comparison constitutes random chance, as the argument classifications for utterances are divided into 4 categories, this initial baseline is 25%.

To construct the BOW, the utterance text minus stop words is used to generate a corpus in order of word frequency, and each tagged utterance is turned into a tuple of (text, tagged reasons). These examples are divided into a train/test split, and trained examples are to generate an X shaped (len(train_examples), len(vocab)) and y shaped (len(train_examples)). For each i in X, a sparse vector with len(vocab) is used, with nonzero values stemming from the count of words from vocabulary utilized in the text from that i, aligned by corpus word ordering. Ex: A [0, 2, 0...] vector would indicate that that training example had 2 instances of the second most popular word in the corpus. Y represents example tags; both numpy arrays are fit to a random forest classifier.

The initial transformer model is XLNet, a generalized autoregressive model that can learn bidirectional contexts through maximizing expected likelihood over permutations, outperforming traditional BERT[5] while integrating ideas from Transformer-XL[7] into pretraining. Utilizing XLNet requires formulation into TFRecords, Google's training formulation for their cloud TPU training. Custom code was written to preprocess the CDD data to TFRecords by turning each utterance into a separate .txt file and creating a folder directory for each respective classification. These folders for 0,1,2, and 2+ were divided into train and test after completion, a custom CDD class was created within the style of the IMDB class within the XLNet models repository for record handling, so that preprocess_classification_data.py might be run to accommodate CDD data. [5]

5 Experiments

5.1 Data

The data provided comes from transcripts from Uganda, US, Malawi, and Ghana. Deliberations were done on coordinated topics at live sites as well as online in multiple groups with a moderator controlling the flow of conversation. These transcripts were then data engineered into csv files with one "utterance' as defined by an uninterrupted piece of speech from a single individual. All human classified training utterances numbered to 11,523 uninterrupted utterances across all locales znd discussion topics. The BOW + RF model train and test data were all held out and separated by topic and location, while as a result of required data points for the XLNet model, train and test data had to be combined, which resulted in a better tested classifier from well diversified data. The task is to classify the number of "reasons" provided in an utterance[4]:

Statement	Reasons
"I support free trade"	0
"Free trade is harmful because it takes jobs away from the US"	1
"“The government should consider universal health care because millions of Americans are uninsured and governments in other countries provide universal health care for their citizens"	2

The test results were a 30% sample of the human coded examples for the BOW model by each subject and a curated selection of TFRecords according to the XLNet script, equally split among the 4 classification possibilities.

5.2 Evaluation method

The evaluation came principally from just how well these systems emulated the results from human coders. The BOW model scores come from the cross-validated average performance of the random forest classifier in classifying against the held out test set for each topic. Conversely, the XLNet model required tuning a custom validation script within the style of run_classifier.py to score the finetuned model as compared to the all category test TFRecords. The nature of the set-up requirements for the techniques necessitated a difference in the nature of train and test data, with the XLNet test data representing a better category (0,1,2,2+) split, and the BOW + RF model allowing for a per discussion metric. In addition, a modified validation script was written to score the CDD data at the general level.

5.3 Experimental details

The RF + BOW was modeled using 1000 a 1000 tree classifier with word embeddings as features. The XLNet model was pretrained bidirectionally with a seq_len 512 and a mask_alpha of 6 on wikidata on a 3.8 TPU. The classifier used the pretrained TFrecords generated through the custom written CDD class within the XLNet model to finetune on the same TPU with a learning rate of 2e-5 with 500 iterations and 4000 trainsteps. The results were then validated against the test set of TFRecords, a representative dataset across the 4 potential CDD argument classification types of 0, 1, 2, 2+.

5.4 Results

Transcript	Random	BOW + RF	XLNet (same for all categories)
Uganda	25%	58%	40%
Malawi	25%	60%	40%
Ghana	25%	55%	40%
US	25%	54%	40%

The qualitative results, at first glance, leave a bit to be desired with respect to the performance of the XLNet model when compared to the BOW + RF method. I don't think this is necessarily all that surprising, given the "harder" test data therein, and I would warrant that despite a lower classification percentage, this is a better generalizable model, and deserving of further finetuning and general exploration. When taken as a "binary" classifier, as in whether a statement has reasons (1,2,2+) or not (0), the transformer model hit accuracy of around 85% on the diversified test set, far greater than the requisite similar metric for any of the RF+BOW technique on any individual test set seen in the graph above.

6 Analysis

As the BOW + RF method operated on a more individual topic basis in a less sophisticated manner, it likely benefited from the large prevalence of "lower" tagged reasons as a guessing or validation tool based on the length of the utterance and the skewed nature of the test data. The classifier may have by proxy more or less just evaluated the length of an utterance and based guess on that, and just generally guessed lower numbers (0 or 1) based on the characteristics of the training data. As the held out test sets for the BOW model were a percentage split of tagged results, the 50-60% baseline seemingly corroborates well with the approximate number of "lower classified" reasons within the training set. This is exemplified by diving further into the Malawi transcript data, with test data split across 20% 0, 60% 1, 14% 2, 6% 2+ within the specific sample. While the model performed 60% in general, it vastly overguessed on the 0,1 side at 95% of model predictions, and only provided a handful of 2+ guesses. Accordingly, if the test set happened to be less diversified and had a very large sampling of 0/1 reason entries, the model would probably do much better and vice versa with a heavier "well-reasoned" test set. Of course, this doesn't necessarily imply a high degree of model viability for real world application. This dynamic serves more as a an explanation for the model's performance on the given test set rather than capacity to emulate CDD argumentation tagging techniques, and a cementing of how difficult it is to get well a volume of well diversified, representative train and test data with live transcript deliberations in each of these well-coordinated, time consuming, and organization intensive live sessions.

7 Conclusion

Counter-intuitively, while the surface level metrics of the transformer based model actually dropped some of the internally examined metrics for ML methods to replicate human coder capacity, I believe they are significantly more applicable to the larger mission. While I have raised concerns about the scarcity of data and in developing a model that uses fewer, not well distributed categorically data points, the response has been that nothing can be done to alter the human tagged samples and training data, and that any sort of attempt at a model to generally structured argumentation was a foolhardy venture. Of course, we are still quite a ways away from any sort of machine level emulation, and the asymptotic performance capacity for the particular CDD method is still quite unknown. Given the slightly amorphous nature of their system. I am proud to say that the exploration into transformers has incredibly promising results in comparison to the initial RF + BOW model I put together based on general requirements for a topic specific basis. I believe that this foray has served as extreme validation, and will portend that once I present the results and analysis into the basis surrounding the minimal difference in accuracy with substantial increase in applicability and further work for the transformer based explorations, there will be a meaningful pivot in research direction. Future work will consist of expanded hyperparameter tuning for CDD data + XLNet, as well as developing general segmentation of transcript data into groups of increasing specificity, both in terms of geographic locale and topic, once data points allow. I am eager to applying the learnings from this class and project to better advance CDD's mission, a goal I wholeheartedly believe in.

References

[1] Amita Misra, Brian Ecker, Marilyn Walker. "Measuring the Similarity of Sentential Arguments in Dialogue". In The 17th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), Los Angeles, California, USA, 2016.

[2] Fishkin, J. S. (1988). Deliberative Polling: Executive Summary. Center for Deliberative Democracy at Stanford University, http://cdd.stanford.edu/polls/docs/summary (accessed February 15, 2020).

[3] Fishkin, J., Garg, N., Gelauff, L., Goel, A., Munagala, K., Sakshuwong, S., ... & Yandamuri, S. (2019). Deliberative Democracy with the Online Deliberation Platform.

[4] Siu, A. (2017). Deliberation & the Challenge of Inequality. Daedalus, 146(3), 119-128.

[5] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." arXiv preprint arXiv:1906.08237 (2019).

[6] Swanson, Reid, et al. "Argument Mining: Extracting Arguments from Online Dialogue." ACL Anthology, Association for Computational Linguistics, 2015, www.aclweb.org/anthology/W15-4631.

[7] Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019.