

Man vs. Machine: Master Debating with Argumentative Dialogue

Stanford CS224N Custom Project

Jason Zhao

Department of Philosophy
Stanford University
jzhao23@cs.stanford.edu

Abstract

Natural language generation involves the machine creation of meaningful text. I address the natural language generation task in an argumentative dialogue setting. Success in this task is important because results can be applied to educational environments as an instance of AI that provokes critical thinking instead of confirming existing preferences. My goal is to create a dialogue system capable of political argumentative conversation with a human user in a sustained message exchange. I present a Transformer-based sequence-to-sequence architecture that generates BERT embeddings from input messages, encodes these embeddings with a Transformer, and then decodes meaningful machine responses through a combination of local and global attention. The Political Language Argumentation Transformer (PLATo) is a novel architecture that achieves lower perplexity and higher accuracy outputs than existing benchmark agents. PLATo surpasses pure RNN and Transformer architectures and provides a generalizable model for future development of argumentative dialogue systems across datasets and use cases.

My mentor is Arnaud. No external collaborators, mentors, or project sharing amongst other courses.

1 Introduction

In our digital era, the explosion of machine learning (ML) has automated many human thought processes, resulting in life-changing technologies. However, proliferation of automation has also entrenched political biases in echo chambers that engender polarization in addition to corporate earnings. I hope that modern ML can be repurposed through the ancient practice of Socratic argumentation towards engaging ideological diversity. Since the ancient Greeks, pedagogy has relied in part on challenging existing beliefs and introducing new insights and frameworks – the very cornerstones of effective argumentative dialogue. As such, this paper uses ML to build a bridge across the political chasm that ML itself has helped create.

Argumentation is a subset of natural language generation (NLG). Recently, natural language processing (NLP) systems have made impressive strides in NLG tasks ranging from question answering on SQUAD [1] to chatbots deployed in customer service applications [2]. Despite widespread advances in language understanding exemplified by these NLG tasks, progress in argumentative dialogue systems (ADS) has been considerably slower. Argumentation imposes additional constraints that are either absent or less prominent in traditional NLG tasks like question answering and application-specific conversation. First, effective ADS require a deep understanding of rhetoric, which is often ambiguous and inconsistent across circumstances. Second, ADS require both global and local contextual awareness throughout a conversation as the logical structure of an argument evolves with time.

Recent literature attempts to resolve these unique challenges using traditional methods. Some ADS

use a retrieval based system that responds to argumentative statements using a database of existing human responses through a similarity metric. These retrieval based solutions produce excellent results when arguments occur in domains native to the dataset, but fail to truly understand rhetoric and adapt to new scenarios. Others attempt to use purely recurrent approaches such as RNN’s and LSTM’s. These are more generalizable than retrieval methods, but struggle to capture a global context in arguments.

I propose a sequence-to-sequence attention-based architecture, PLATo (Political Language Argumentation Transformer). PLATo uses BERT [3] to generate embeddings that are then encoded through a **local-level encoder** implemented as a Transformer [4]. The encoding produced by the local-level encoder is then attended to by an attention distribution generated by a **global-level** memory implemented as an LSTM. The hidden state stores previous responses from the conversation thus far and acts as global context. Finally, the attended encoding is passed to a **local-level decoder** implemented as a Transformer, which produces the machine response. PLATo achieves lower perplexity, higher accuracy, and more coherent responses than existing ADS. These results will be covered extensively later in the paper. In the following related works section, I will explain the advantages of PLATo over legacy solutions through a survey of related works upon which I expand.

2 Related Work

The vast majority of NLP research in argumentation revolves around identifying claims [5][6], classifying stances [7][8], and labeling political bias[9]. In contrast, *generating* reasonable arguments is a relatively nascent field. In this section, I will track the historical development of ADS as a subfield of NLG.

2.1 Retrieval Systems

Retrieval systems *select* a relevant response from a stored database given a user message. In order to select the best response, retrieval systems typically use some similarity function between the user message and stored responses in the database. These similarity functions range from LSTM networks in [10] to Latent Semantic Analysis, a statistical method for determining similarity, in [11]. Retrieval works well in pre-defined topic settings. However, when the user’s argumentation becomes more complex or moves out of the domain of existing datasets, retrieval systems cannot generate novel responses that do not already exist in the database.

2.2 Generative Systems

In response to the limitations of retrieval systems, generative systems can compose new messages based on the context of the argumentation. Early attempts to do so, such as [12], leveraged the sequence-to-sequence encoder and decoder architecture – proposed by [13] – with promising results. These architectures typically used RNNs or LSTMs to encode input, and then this encoding would be decoded by another RNN or LSTM to produce machine output. While these early generative sequence to sequence systems were more adaptive to diverse circumstances and often produced more semantically robust responses than retrieval systems, they still struggled with keeping track of the global context of an argument. As such, irrelevant or repeated responses were frequent. [12]

2.2.1 Global Session-Level Memory

In response to the lack of global memory, [14] presents a novel session-level memory that attempts to capture the entire context of a conversation. Thus, the model does not greedily respond to only the most recent user input, but also considers the arc of the entire conversation. This session-level memory is implemented as an RNN, the output of which is softmaxed into a distribution that attends to the encoding output by the encoder. Since the encoder only receives the most recent user input, this attention helps re-weight the encoding with knowledge from the global session-level memory.

2.2.2 Self-Attention and Transformers

With the rise of self-attention and its ability to outperform recurrent models, Transformer-based architectures have become the new norm in NLP. We see several ADS, such as [15] and [16], that

have built off of the global attention method developed in [14] by substituting transformers instead of LSTM's for the encoder and decoder. This has resulted in a slight improvement in perplexity, though general issues in repeating ambiguous phrases (e.g. "I don't know") still hamper fluency.

2.2.3 PLATo and BERT Embeddings

The current leading NLP architecture is BERT (Bidirectional Encoder Representations from Transformers).[3] With its contextual word embeddings, BERT has become the general paradigm for NLP. I want to integrate the BERT embeddings with existing advances outlined above. By doing so, I hope to improve perplexity and subjective fluency by producing more nuanced and context-aware responses to prevent the repetitive and irrelevant responses that frustrate even the best ADS. Thus, PLATo is a generative, sequence-to-sequence, transformer-based argumentative dialogue system with global memory that uses BERT embeddings to achieve better results.

3 Approach

3.1 Baselines

As my baseline, I use the perplexity scores from "Dave the Debater" [14], the most prominent argumentative dialogue agent. The perplexity metric used in [14] that I will benchmark against can be defined over examples y (where $P(y)$ is the log-likelihood) as:

$$PP = 2^{-\sum_y \log P(y)}$$

Because "Dave the Debater" is the most well known debate agent and uses the same dataset as my approach (Internet Argument Corpus, see 4.1), using the published perplexity in [14] provides a relevant comparison. "Dave the Debater" achieved a perplexity score of **88.51** that I will seek to surpass. While other methods in the literature reviewed above, particularly [16], have achieved modest perplexity gains over "Dave the Debater", they do not offer as much actual machine-generated text to compare subjective fluency against and are not as frequently cited.

3.2 Architecture

1. **BERT Embedding:** The input is a user-input response r^m , which is the most recent addition to the set of responses $\mathcal{R} = r^1, r^2, \dots, r^m$ where m is the number of responses in the current session. Each response is represented by words $r^m = w_1^m, w_2^m, \dots, w_n^m$, where n is the total number of words in response m . Each response r^m is input into a BERT-base model [3] that returns the contextual word embedding for each individual word w_i^m in r^m . These word embeddings are of dimension 768. A linear transformation then projects each word embedding into a smaller size d_{enc} vector.

$$\textbf{Embedding: } \mathbb{R}^n \rightarrow \mathbb{R}^{n \times 768}$$

$$\textbf{Linear: } \mathbb{R}^{n \times 768} \rightarrow \mathbb{R}^{n \times d_{enc}}$$

2. **Encoder:** A **local-level encoder** is implemented as a transformer with multi-headed self-attention and position-wise feed-forward layers. It reads in the embedding for each word w_n^m and returns its encoding; I max-pool over the encoder output for the response r^m ($\mathbb{R}^{n \times d_{enc}}$) to receive r_{enc}^m ($\mathbb{R}^{d_{enc}}$).

$$\textbf{Encoding: } \mathbb{R}^{n \times d_{enc}} \rightarrow \mathbb{R}^{n \times d_{enc}}$$

$$\textbf{Max-Pool: } \mathbb{R}^{n \times d_{enc}} \rightarrow \mathbb{R}^{d_{enc}}$$

3. **Global Memory:** Next, a **global-level memory**, implemented as an LSTM with hidden size d_h , takes in the max-pooled encoding r_{enc}^m produced by the local-level encoder. I attend to the local-level encoder's output by softmaxing the LSTM generated distribution and weighting each position in the local-level encoder's output with its attention weight. This global-level memory helps the model remember the overall information exchange in the argument by storing the context of all responses the model has seen in the LSTM hidden state.

$$\textbf{LSTM: } \mathbb{R}^{d_{enc}} \rightarrow \mathbb{R}^{d_h}$$

$$\textbf{Attention: } \mathbb{R}^{d_h}, \mathbb{R}^{n \times d_{enc}} \rightarrow \mathbb{R}^{n \times d_{enc}}$$

4 Experiments

4.1 Data

I use the Internet Argument Corpus (IAC) dataset [17], used for research in *political* debate over internet forums. IAC includes around 11k discussions, 390k messages, and 73m words. Python preprocessing code is included that manages the extraction of messages from discussions. IAC will be used to learn the task of generating a reasonable argumentative response: for every message r_i in IAC, the machine will attempt to generate the appropriate response r_{i+1} to that message.

4.2 Evaluation Methods

The objective function during training is MMI loss (see 3.2). Evaluation is done through perplexity (see 3.1) in order to compare with the baseline [14]. An auxiliary metric word accuracy, the exact token by token match between target and output sequences. The target sequence is the actual human response r_{i+1} to the input r_i , and the output sequence is the response generated by PLATo to r_i .

4.3 Experimental Details

These hyperparameters are optimized for a model configured according to 3.2. I used the BERT-base model across all experiments. The following settings were used to obtain all results presented in section 4.4. **Learning rate** was found to work best at $1e-3$, with $1e-2$ and $1e-4$ performing significantly worse. **Dropout** is set to 0.1 for both encoder and decoder. **Embedding size** was set at $d_{enc} = 300$ to balance efficiency and representational power; **LSTM hidden size** was set at $d_h = 512$. **Adam optimizer** was used with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. **Training duration** was 5 epochs on all experiments, and each epoch required approximately 10 hours to complete.

4.4 Results

4.4.1 Best Performance Against Baseline

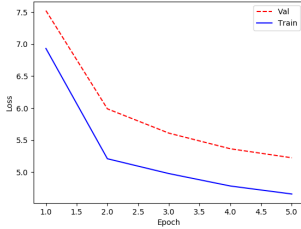


Fig. 1: PLATo-MLE Loss

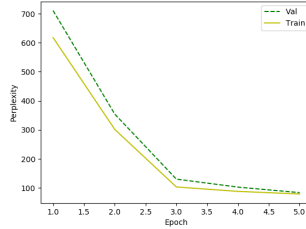


Fig. 2: PLATo-MLE Perplexity

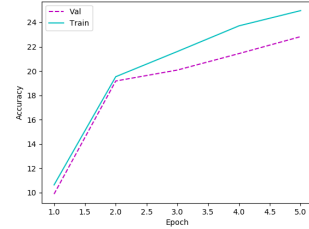


Fig. 3: PLATo-MLE Accuracy

The best performance was achieved with the exact model configuration outlined in 3.2, with one important change: Maximum Likelihood Estimate (MLE) loss was substituted in for Maximum Mutual Information (MMI) loss for better results. The MLE loss for target response r^* is defined as:

$$r^* = \operatorname{argmax}_r \log P(r|r^{1:m})$$

This optimal configuration, which I call PLATo-MLE, achieved validation perplexity of **83.69** (Fig. 2) after 5 epochs, which exceeded the validation perplexity of "Dave the Debater" of **88.51** after over 40 epochs. The loss and accuracy curves reflect the perplexity improvement over "Dave the Debater" (Fig 1, 3). This corresponds with current state-of-the-art in ADS and could improve given further training given the loss curve. Below are some comparisons between "Dave the Debater" and PLATo that highlight the subjectively similar qualities of their outputs.

Query	Dave	PLATo-MLE
Sure, I believe in God.	I am not a christian but there are so many of the bible.	I'm not sure that you are a Christian. I don't think that the bible is a Christian.

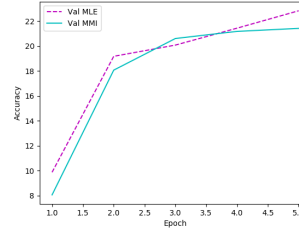
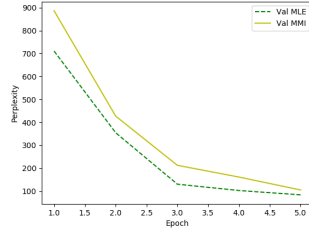


Fig. 4: MMI vs MLE Perplexity Fig. 5: MMI vs MLE Accuracy

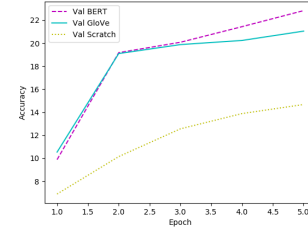
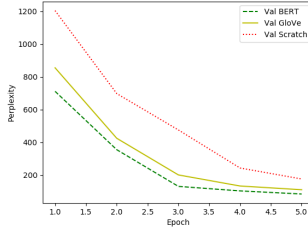
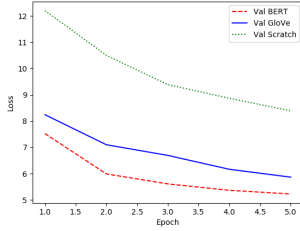


Fig. 6: Embeddings Loss

Fig. 7: Embeddings Perplexity

Fig. 8: Embeddings Accuracy

4.4.2 MMI Loss vs MLE Loss

As mentioned in 3.2, MMI loss achieved superior results to MLE loss for "Dave the Debater". [14] However, I discover that PLATo achieves superior performance with MLE over MMI. The validation perplexity of PLATo-MLE (the configuration in 3.2 except with MLE instead of MMI loss) was **83.69** and that of PLATo-MMI (the exact configuration in 3.2) was **104.69** (Fig. 4). The validation accuracy confirms the superiority of PLATo-MLE (Fig. 5). While this result is unexpected and may converge differently with a longer training time, I offer an explanation in 5.1.

4.4.3 Embeddings

Different embedding options produced predicted results (Fig. 6-8). All embedding experiments were performed on PLATo-MLE. I test with training embeddings from scratch, pre-trained GloVe embeddings, and contextual BERT embeddings. Note that I did not fine-tune the BERT model due to time constraints; the model is a pre-trained BERT-base. The embeddings trained from scratch produced poor results as the IAC dataset is neither large nor representative enough to result in robust word embeddings, especially through only 5 epochs. The GloVe embeddings yielded a considerable performance increase, and the BERT embeddings added a moderate performance boost over GloVe.

5 Analysis

5.1 Analysis of Loss Functions

Query	PLATo-MLE	PLATo-MMI
Complete restriction is the best strategy for gun control.	I don't think that the government is the government.	I don't think that the US is quaida quaida.

It is surprising that MLE outperformed MMI as a loss function, because MMI loss is designed to prioritize diversity in generated responses by penalizing common words, and was used with greater success than MLE in "Dave the Debater". [14] However, upon further qualitative analysis, I believe that MLE outperforms MMI in PLATo for two reasons. First, [14] does not use BERT or GloVe embeddings, instead opting for embeddings trained from scratch. Embeddings trained solely on IAC tend to be semantically deficient and produce long sequences of common words (e.g. "a", "the", and "I"; see 5.2). Because of this, MMI is helpful to [14] in generating more diverse responses. However, because PLATo harnesses robust contextual understanding from BERT embeddings, the diversity

incentive that MMI imposes actually forces PLATo-MMI to generate less common words ("quaida") that make little sense in the context of the response, whereas PLATo-MLE does not suffer from this issue, producing words that are semantically relevant to the query ("government").

5.2 Analysis of Embedding Method

Query	BERT Embed	GloVe Embed	Trained Embed
I believe that God exists because the Bible tells the truth.	I don't think that the bible is christian. I do not have to.	I don't know that you are not believe that god is not god.	the the the the the the the the the the the

The superior performance of BERT embeddings over GloVe, and GloVe over training embeddings from scratch, is an expected and well-understood phenomenon. BERT embeddings provide powerful context awareness, GloVe vectors do not have this context awareness but, like BERT, have been pre-trained on large datasets, and training from scratch has none of these advantages. Thus, the quantitative results require no further explanation; instead, I analyze the specific qualitative distinctions in PLATo-MLE's output between each embedding strategy. First, it is obvious that embeddings trained from scratch simply do not understand the semantic and syntactical relationship between words well enough to replicate any semblance of grammar. The IAC is neither large nor general enough in its contents to produce robust embeddings, especially given only 5 epochs of training. Second, GloVe embeddings clearly show domain knowledge. When prompted about believing God and the Bible, PLATo-MLE responds with a relevant (if meaningless) phrase: "...believe that god is not god". At the least, PLATo-MLE can produce on-topic responses with GloVe embeddings. With BERT embeddings, PLATo-MLE is not only on-topic, but also offers some meaningful interpretation. Further, its response introduces the word "christian", a related word not included in the query.

5.3 Analysis of Training Progression

Query	Epoch 1	Epoch 3	Epoch 5
Do you really think God exists? If so, why do you think so?	I do not, and I have, and I, I have	I don't think that the bible is christian.	I don't see the bible, but christian do not have to.

We see that PLATo's lower quantitative perplexity score is comprised of qualitative improvements in both *relevance* and *responsiveness*. Relevance refers to the aptitude of the words in PLATo's response with respect to the query. Responsiveness refers to the actual semantic meaning of PLATo's response with respect to the query. For example, a high relevance, low responsiveness response to the question "Do you believe in God?" would be "God", as it includes a relevant word but does not answer the question in any way. In contrast, a low relevance, high responsiveness response would be "No", for the opposite reason. With more epochs, PLATo generates more relevant words and the semantic meanings of its responses better address the question; for the most part, PLATo is neither only learning to match words nor only learning to repeat the same (meaningful, yet not very relevant) phrase (for example, "I don't know").

However, an important qualification is that PLATo, while it does not tend to default to responsive yet irrelevant phrases such as "I'm not sure" or "Yes", does typically generate sentences with similar structures, almost always including some variation of "I do/don't think" or "I am/am not sure". I believe this is largely a result of the dataset limitations discussed in the next section.

5.4 Dataset Limitations

5.4.1 Size

Query	PLATo-MLE
Clearly your argument are jokes.	I'm don't think I'm sure.

The Internet Argument Corpus (IAC) [17] hampers PLATo's general performance in two ways. First, it is not large enough. While the 390k messages are certainly enough to learn basic dialogue, truly meaningful argumentation requires much more data. Sarcasm, rhetoric, and metaphors are all

complex and highly personal elements of the human language that will require at least an order of magnitude more argumentative data than is currently available. We see as an example that PLATo will respond to direct insults with a serious and oblivious tone.

5.4.2 Coverage

Query	PLATo-MLE
School shooters are evil. We can all agree on that. We need to prevent them.	I'm not sure that the Bible is a good problem, but I don't believe in God.

Second, the IAC is very limited in its domains. While generative dialogue models are typically better at adapting to new topics than retrieval-based systems, they still struggle on domains of conversation that are data-sparse. PLATo is quite effective in responding to claims or questions about Christianity, but far worse at discussing education, taxes, or moral claims about good and evil. We see that PLATo often even uses Christian terminology even when it is completely uncalled for, such as in a question on "school shootings".

5.4.3 Learned Bias

Query	PLATo-MLE
You support gun abuse, then.	I'm not sure that you are not a gay sex.
Black people deserve justice from racism and oppression from whites.	I'm don't think that they are the same.

Aside from the general limitations of the IAC, perhaps the most concerning aspect of PLATo's behavior is its incredibly racist and sexist behavior. Unfortunately, to my knowledge, *none* of the literature on argumentative dialogue systems have dealt with machine bias, even as the general field of NLG has moved towards the pressing and salient challenge of machine ethics.

I identify two types of bias, stemming from two different causes. First, *active* bias learned from the dataset. This is a bias that PLATo actively develops because of the unethical and unsavory dialogue contained within the forums that the IAC is comprised of. Because these are unregulated online forums, and because the IAC has done little to filter these forums, there are many examples of nonsensical and harmful correlations that PLATo then replicates, such as the relationship between "gun abuse" and "gay sex". As with all of the qualitative examples shown thus far, this is not a cherry-picked example; it is reflective of a *general* association in BERT's ontology.

Second, *passive* bias from lack of semantic understanding. This is an passive bias that PLATo obtains because of its ignorance of the query's meaning as opposed to something actively learned from the dataset. When PLATo does not understand a query, it often produces some variance of "I don't think that..." or "I'm not sure..." In unfortunate circumstances, this reflects as an accidental, yet very real, bias as in the example shown regarding racial relations.

6 Conclusion

I present a novel architecture, the Political Language Argumentation Transformer (PLATo). PLATo integrates existing advances across NLP to create a state-of-the-art architecture: a sequence-to-sequence model that utilizes Transformers, global-attention, and BERT embeddings. PLATo achieves quantitative performance gains in perplexity over existing argumentative dialogue systems. Qualitative analysis suggests that MLE loss is superior to MMI loss given more robust word embeddings, and that BERT embeddings are the most effective embeddings for ADS. Further, PLATo learns both *responsive* and *relevant* arguments over time. Despite these advantages, the IAC dataset is inherently limited by size and coverage, and much more attention needs to be paid to the ethical ramifications of both *active* and *passive* bias learned by ADS. Finally, this paper does not analyze the contextual awareness of PLATo's responses with respect to the history of a longer conversation because PLATo (and ADS) in general are not robust enough to consistently provide *singular* meaningful argumentative responses, much less an entire conversation's worth. Future work should focus on generalizing PLATo to larger datasets, mitigating machine bias, training for longer periods of time, and ultimately attempting to move ADS towards longer conversations by first consistently mastering one-off responses.

References

- [1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.
- [2] Pranav Fadhil and Gianluca Schiavo. Designing for health chatbots. In *Pre-print*, 2019.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- [4] Ashish Vaswani, Noam Shazeer, and Niki Parmar. Attention is all you need. In *NIPS*, 2017.
- [5] Filip Boltuzic and Jan Snajder. Identifying prominent arguments in online debates using semantic textual similarity. In *First Workshop on Argument Mining*, 2015.
- [6] Ran Levy, Yonatan Bilu, and Daniel Hershcovich. Context dependent claim detection. 2014.
- [7] Kazi Hasan and Vincent Ng. Stance classification of ideological debates: Data, models, features, and constraints. In *Joint Conference on Natural Language Processing*, 2013.
- [8] Mohit Iyyer, Peter Enns, and Philip Resnik. Political ideology detection using recursive neural networks. In *Association for Computational Linguistics*, 2014.
- [9] Noah Smith and David Bamman. Open extraction of fine-grained political statements. In *Conference on Empirical Methods in Natural Language Processing*, 2015.
- [10] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on International Conference on Machine Learning*, 2014.
- [11] Jason Williams, Kavosh Asadi, and Geoffrey Zweig. Hybrid code networks: practical and efficient. In *CoRR*, 2017.
- [12] Quoc Le and Oriol Vinyals. A neural conversational model. In *CoRR*, 2015.
- [13] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.
- [14] Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Association for Computational Linguistics (ACL)*, 2018.
- [15] Ademi Adeniji, Nate Lee, and Vincent Liu. Sequence-to-sequence generative argumentative dialogue systems with self-attention. In *Stanford University*, 2019.
- [16] Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn Walker. Debbie, the debate bot of the future. In *Springer International Publishing*, 2019.
- [17] Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, 2012.

A Appendix

Query	PLATo-MLE
Should the author receive a stellar grade on this paper?	I’m not sure that the two are the same.