Does Classifying Principled Arguments Improve Understanding of Stance?

Purdue CS 577 Project

Joshua Yeung
Department of Computer Science
Purdue University
yeungj@purdue.edu

Albert Xu
Department of Computer Science
Purdue University
xu1018@purdue.edu

Minh Nguyen
Department of Computer Science
Purdue University
nguye330@purdue.edu

Daniel Hu
Department of Computer Science
Purdue University
hu432@purdue.edu

Abstract

Debate is a challenging task for computers and competitive debaters alike. It involves responding relevantly and argumentatively to dialogue, in addition to the challenge of generating coherent counterarguments. Argument mining employs many techniques and methods, one of which involves the idea of principled arguments. This project attempts to answer questions about the utility of principled arguments in the context of debate. We test this by training several models to predict argument stance from claim. Our results show that the classifiers using principled argument information outperform the classifiers not using the information, so we claim that principled arguments do improve understanding of debate.

1 Introduction

Argument mining is a relatively new research field pursuing highly challenging tasks dealing with automatic extraction and identification of argumentative structures from natural language. These structures include premises, conclusions, argument schemes and the relationship between pairs of arguments and their components. 'Argument Invention from First Principles' [1] proposes the existence of principled recurring arguments which are relevant to many topics and defines a taxonomy of such arguments. The paper aims to automatically identify classes of relevant principled arguments (from a manually created set of principled arguments) given a motion. The authors achieve a reasonable amount of success with their motion-argument matching, and we would like to take the authors' work a step further.

Our main task for this project is to attempt and identify claim stance based on the principled arguments the claim is most similar to. We want to compare the performance of a stance classifier that uses only the input sentence versus a classifier that uses knowledge of principle arguments.

Currently, the idea of principled arguments, from the 'Argument Invention...' paper, are used with the intention of creating new arguments from a topic. However, the concept is novel and unproven. We attempt to bolster the validity of the idea of principled arguments by showing that their knowledge is essential in understanding debate; that their knowledge can improve the performance of a stance classifier.

1.1 Novel Aspects

Our novel contribution to stance classification is the addition of knowledge of principled arguments. Specifically, we are trying to identify a specific concept that is vital to understanding debate, as well as providing a novel method of integrating that knowledge into the algorithm.

2 Problem Definition

Our project is a classification problem. Given a sentence s, predict whether it is pro- or con-. i.e. predict one of two labels. Each sentence s is associated with a vector embedding v, through some transform $E(s) = v$. The principle arguments of interest are also sentences, for example "adolescents are as capable as adults." We will refer to the 72 principle arguments as p_i, with $E(p_i) = \vec{a}_i$

The objective of our project is to show that there is some model M such that the probability our model predicts the correct label $P_{M(s,p_i)}$ (using the principle arguments) is greater than the probability without using the principle arguments $P_{M(s)}$. That is, show $P_{M(s,p_i)} > P_{M(s)}$.

3 Technical Approach

Our technical approach involves comparing many different methods. From our problem definition, we have the following 3 functions to define.

$$E(s) : s \to v$$
$$T(v, p_i) : v, p_i \to v$$
$$M(v) : v \to \{\text{pro}, \text{con}\}$$

$E(s)$ is the embedding function, turning a sentence into a machine-usable vector of numbers. $M(v)$ is the classification step, where we use the embedding to predict an output. $T(v, p_i)$ is our novel aspect, where we try to incorporate principle argument information.

We use three types of embedding functions: learned embeddings and pre-trained embeddings (averaged Word2Vec, TF-IDF on W2V, SBERT, and Facebook's Infersent encoder [2]). We will compare the embedding methods with each other on downstream task performance (stance prediction).

Sentence-BERT (SBERT), a modification of the original pretrained BERT network (Devlin et al., 2018)[3], is fine-tuned with siamese and triplet network structures to derive semantically meaningful sentence embeddings which can be compared using cosine-similarity (Reimers et al., 2018)[4]. This method should work better than our Word2Vec-based methods because BERT produces contextualized embeddings through its bidirectional transformer architecture, while the other two rely on context insensitive word embeddings. We chose SBERT to maintain the accuracy from BERT while reducing its overhead [4]. Facebook's Infersent encoder uses a BiLSTM architecture with max pooling. It was trained on Stanford's Natural Language Inference dataset using GloVe[5] vectors as base features. Theoretically, this method is similar to BERT as it captures context in its embeddings by using a BiLSTM.

Our classification $M(v)$ will be a simple logistic regression with a single linear layer. The following sections are about choices for $T(v, p_i)$, our primary novel contribution.

3.1 Baseline

The baseline is to not use any information about the principled arguments. We train a classifier that directly predicts stance from a claim.

$$T(v, p_i) = v$$

3.2 Principle Argument Distribution, using Attention

In this architecture, we attempt to model the probability distribution of the stance of a claim conditioned on a number of principle arguments. To this end, we encode the claim into a vector $\vec{v}$, and all principle arguments into vectors $\vec{p_i}$. Using an attention mechanism inspired by the mechanism proposed by Bahdanau et al. [6], we feed the encoded claim and principle arguments to the mechanism. We take the inner product of encoded principle arguments with the claim projected using an attention matrix A, and feed the result to a softmax layer, which outputs the weights of each principle arguments.

$$\begin{gathered}
\alpha_i = \frac{\vec{p}_i}{|\vec{p}_i|} \cdot A\vec{v} \\
\vec{\beta} = softmax(\vec{\alpha}) \\
\vec{z} = \sum_i \beta_i \vec{p}_i \\
T(v, p_i) = \vec{z}
\end{gathered}$$

3.3 Principle Argument Score Decoder

This approach uses the same projected score $\vec{\beta}$ as the previous section. However, rather than taking the weighted sum of the principled argument embeddings, we interpret the prior formulation as a single-layer encoder that is based primarily on the principled argument embeddings. Then we can treat this as a translation problem, using a single layer decoder D to project the sentence claim into a vector encoding both the claim and the principled-arguments, where W is the weight matrix of a single-layer decoder.

$$T(v, p_i) = D[\beta] = \text{LeakyReLU}(W\vec{\beta})$$

4 Evaluation

4.1 Rationale

We compare the model's performance based on their accuracy in classifying the stances, a binary classification problem. We claim that if knowledge of principled arguments improve understanding of debate, then the performance of models that use principled argument knowledge will be better than the baseline model that does not carry this information.

We're also interested in comparing our methods of incorporating principled argument information, as it's not well-explored and there is no standard way to incorporate the information yet.

4.2 Experimental Settings

The data is tested on the reasons dataset from UT Dallas [7]. We will only use stance field from their data, ignoring their nuanced stance fields, like 'con-baby rights' for the abortion topic.

We will run 6 experiments, comparing our 2 models and a baseline, using 2 different embedding methods. We learned our own embeddings and tested them against pretrained embeddings. Each of the embedding and T-function pairs listed below will be evaluated using k-fold cross validation (k=5).

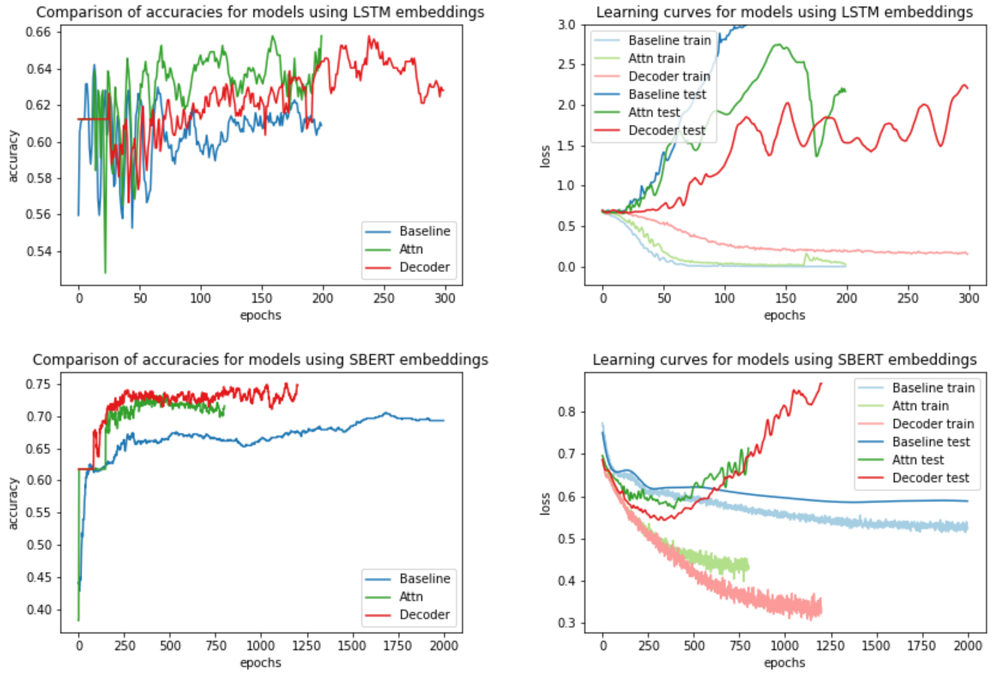
Embedding	T function	Learning rate	Dropout rate
BiLSTM	Baseline	1e-4	0.3
BiLSTM	Attn	1e-4	0.3
BiLSTM	Decoder	1e-4	0.3
SBERT	Baseline	5e-6	0.5
SBERT	Attn	5e-6	0.5
SBERT	Decoder	5e-6	0.5

4.3 Results

The results of our k-fold cross validation are summarized below. Notably, the maxes are quite different from the mean, and all come from the same fold, with the exception of 'Learned + Baseline'.

Model	Mean accuracy	Max accuracy
BiLSTM + Baseline	0.6450	0.7035
BiLSTM + Attn	0.6468	0.7329
BiLSTM + Decoder	0.6675	0.7381
SBERT + Baseline	0.6865	0.7575
SBERT + Attn	0.6942	0.7592
SBERT + Decoder	0.7103	0.7469

Our models consistently outperformed the baselines with both the BiLSTM and SBERT sentence encoders. The models trained using SBERT encodings performed better across the board than the models trained on any other embeddings/encodings. Additionally, looking at the learning curve of one fold, we can tell that the models using principled argument information have a higher learning capacity than the baseline.



(a) Plot of accuracy versus epoch number. Here we can see that both models using principled argument information outperform the baseline model.

(b) Plot of the learning curves of each used model. We can see both models using principled argument information start overfitting rather early.

5 Summary

In summary, the models using principled arguments outperformed the baseline by several percentage points in all cases. This suggests that our hypothesis is true, that knowledge of principled arguments does in fact improve understanding of debate text.

If we had more time, we would attempt to run this classification on more datasets with more diverse topics, since the dataset we trained on only had claims/arguments for 4 topics: Obama, abortion. marijuana, and gay rights. We could also group datasets by debate topic and see if the principle arguments associated with a particular topic serve as useful features to aid with classification.

References

[1] Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[7] Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, 2014.