Evidence Quality: IBM Debater

Satenik Rafayelyan
Deep Learning for Natural Language Processing – SS20 – Philipp Cimiano and Philipp Heinisch

June 25, 2023

1 Introduction

As a rule of thumb most interesting questions do not have an exact answer, rather there do exist arguments which support or reject them. When there is a debatable topic each side tries to prove their point by bringing the most convincing pieces of evidence. But what makes an argument more persuasive? In this paper the problem of assessing evidence convincingness is described, which is considered a new challenging task in computational argumentation.

Measuring qualitative features of arguments, particularly convincingness, has started to gain more attention in the field of computational argumentation. Essay persuasiveness scoring is one of the examples of the application areas which has been represented by (Ghosh et al., 2016), where authors investigate whether argumentation features derived from a coarse-grained argumentative structure of essays can help predict essays scores, on the other hand (Farra, Somasundaran, and Burstein, 2015) focus on methods for identifying opinion expressions rather then investigating the argument structure. Another field of ongoing interest is social media. Since everyone including politicians and voters today increasingly turn to social media to attract others to their cause, (Hidey and McKeown, 2018) try to identify when a post will be influential or convincing which can help to understand the appeal of political candidates and the reaction to current issues. It is worth mentioning that the authors consider the problem of persuasion prediction a difficult task as it requires modeling world knowledge, social interaction, and reasoning.

Considering the growing interest and beneficial impact on the area, IBM research team (Gleize et al., 2019) released IBM-EviConv data set, which is the topic of analysis of this paper. The data set contains evidence pairs offering a more focused view of the argument convincingness task. As a source of evidence sentences they use the evidence data set released by (Shnarch et al., 2018), which contains more than 2,000 evidence sentences over 118 topics, afterwards they sampled more than 8,000 pairs of evidence and sent them for manual convincingness labeling. Overall the final data set consists of 1,884 unique evidence sentences which spread over almost 70 different debatable topics. For each topic evidence pairs are on the same side, arguing either for the topic (PRO) or contesting it (CON). After cleaning step only 5,697 pairs are left (4,319 train data and 1,378 test data). The detailed description of the data set, such as prepossessing, labeling procedure and so on, can be found in the original paper.

In this work solve the evidence convincingness classification task by using pre-trained language models, particularly Bidirectional Encoder Representations from Transformers (BERT). There currently exist other similar data sets and papers which try to appropriately classify the more convincing evidences. This is discussed more detailed in the next chapter "Related work". In the "Method" section the approach to the problem is represented with a comprehensive description of the model architecture and selected language model. Afterwards, the "Evaluation" part provides the test results and compares them to the results received in the related work. And finally, the "Conclusion" section draws a conclusion and summarizes the work.

2 Related work

There exist several previous works regarding the argument quality assessments and particularly focusing on the convincingness of the arguments. One of the most discussed papers by (Habernal and

Gurevych, 2016) represents empirical assessment of reasons for argument convincingness using Bidirectional LSTM network and SVM applied on UKPConvArg2 data set which the authors released directly with the research paper. Another work that addresses persuasiveness of the arguments has been discussed by (Gleize et al., 2019) mentioned also in the introduction part of this paper. Because the data set which is used here coincides with the data set released and analyzed by the authors, next we will summarize their methodology and approach to the problem.

In order to identify the most convincing arguments in a set it is proposed to rely on the ideas from the field of learning to rank. Ranking can be formalized in various ways; 1) in a pointwise approach, the input consists of a single element and for each input the output is a score, therefore in this case ranking is performed by simply ordering the elements by their scores, 2) in a pairwise approach the input consists of pairs of elements and the output corresponding to each pair of input is the preference between the input elements, consequently in this case ranking means comparing all pair combinations. The paper discussed in the begining of this section based on BiLSTM and SVM is able to provide pairwise inference only meaning that they cannot predict a convincingness score for a single argument. Considering this disadvantage the authors of the paper propose Siamese network structure achieving competitive results on the task of argument convincingness classification and ranking.

Siamese network consists of two legs of identical networks, which share all their parameters and are connected at the top with a softmax, the output of their proposed network is a probability. Each leg in the Siamese network is a neural network which is a function of an input of one argument and has two outputs which show how convincing the argument is and a dummy output which can be a constant. During the training the softmax function is applied to the output of two legs representing the covincingness of each input argument and then the result is compared to the label of the pair using the cross entropy classification loss. Similarly, to predict the convincingness score for only one argument only one leg need to be fed, and then the softmax can be applied to the convincingness output and the untrained dummy output. Each leg in the Siamese network is a BiLSTM which are fed with non-trainable word2vec embeddings followed by 100 attention heads and a fully connected layer with two outputs. Training was done using Adam optimizer, applying gradient clipping and dropout. As a result the network is able to achieve 73 % accuracy on the IBM-EviConv data set.

As it is suggested in the paper there is a room for improvement which can be achieved by pre-training one leg of the network on an argument detection data set. Although the network structure used in this paper differs from the one that they used, but the suggestion of using pre-trained language models is adopted along with the output structure, meaning that the developed network here outputs probabilities.

3 Method

The method developed here is based upon a powerful language representational model named Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). BERT has been extensively trained over large corpus of data to perform two tasks:

1. Masked Language Model - predict the missing word by randomly replacing words with a predefined token, [MASK]

2. Next Sentence Prediction - given a pair of sentences predict whether one follows the other

Because of its bidirectional nature BERT models analyze every sentence with no specific direction, namely the model can understand the meaning of each word based on the context both to the right and to the left of the word, obviously giving the model a clear advantage in the field of context learning. Being pre-trained on 2,500 million internet words and 800 million words of Book Corpus, BERT is able to achieve state of the art results and because of its stronger awareness of the context, usually outperforms other language models.

Since BERT is very big and even fine-tuning takes lots of time, "bert-base-uncased" size variant is used in the model for a binary classification task. The fine-tuning process is initialized with weights from the general purpose pre-trained model but task specific trainable layers are also added on top of frozen layers to adapt the pre-trained features on the IBM-EviConv data. Following standard practice

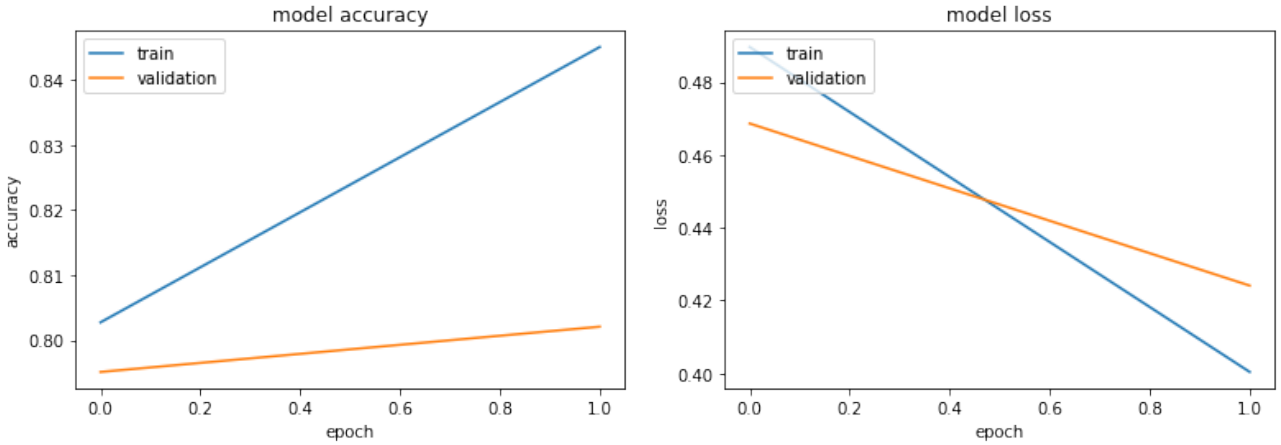

Figure 1: Left: accuracy of the model on the train and validation datasets during the training. Right: loss of the model on the train and validation datasets during the training

with BERT given a pair of evidences first of all BERT tokenizer is applied, meaning that input ids, attention masks and token type ids are derived, which are used as input by the model. Input ids are token indices, numerical representations of tokens building the sequences. Since we use two sequences which do not necessarily have the same length, padding can be applied to the shorter sentence and attention masks take care of adding special tokens when using padding, so that the attention layers will ignore that special tokens. And finally, the token type ids serve for separating the sequences; given a pair of evidences A and B, the network is fed with the following sequence $[CLS]A[SEP]B$; the $[CLS]$ token indicates the contextual embedding for the whole sequence wheres the $[SEP]$ token shows that the input should be treated as a pair.

After loading pre-trained BERT model we freeze it to reuse the pre-trained features without modifying them. As already explained then the above mentioned three findings are used to build the BERT model. Afterwards, trainable layers are added on top of frozen layers to adapt the pre-trained features on the IBM-EviConv data. Concretely, bidirectional LSTM (BiLSTM) with 64 units is added and then hybrid pooling (consisting of average pooling and max pooling) is applied to the output of BiLSTM. Dropout rate of 0.3 is used in order to minimize the overfitting during the training. As well as softmax activation function is applied to the last layer in order to make the model output probabilities of two possible classes. Training is done using Adam optimizer. After training with pre-trained BERT model, fine tuning is applied by unfreezing the BERT model and retraining it with a very low learning rate, in this case 10^{-5}. This delivers meaningful improvement by incrementally adapting the pre-trained features to the IBM-EviConv data.

In addition to the proposed architecture which gives the best accuracy, different experiments have also been conducted in terms of applying different layers and parameters. The results of the experiments as well as the best outcomes with corresponding plots are shown in the next section.

4 Evaluation

In this section we evaluate the methods described in the previous chapters. Various experiments are conducted and corresponding results are reported in the original paper (Gleize et al., 2019), so for the sake of completeness those results are presented along with ours. In order to compare the outcomes of different methods the accuracy metrics are documented in Table 1. In addition to the table we report also the accuracy and the loss of train and validation data sets during the training. See Figure 1.

As it is obvious from Table 1 our method based on BERT outperforms the previous results. It is worth mentioning that there is not a significant sign of overfitting, because it would happen if model's loss on the training set is very low but then, it is large enough on the test set.

It has been mentioned in the previous chapter that different experiments have been performed

Model	Accuracy
Evidence length	0.53
Most frequent label	0.54
Detection model	0.59
GPPL	0.67
GPPL opt.	0.67
GPC	0.67
EviConvNet	0.73
Based on BERT	**0.75**

Table 1: Accuracy on IBM-EviConv. our model (Based on BERT) outperforms all prior results

around the developed BERT-based model. Maximum length of the input sequence has been sequentially decreased but it increased the training time without giving a valuable accuracy improvement. Additionally, trainable dense layers have been added on top of the frozen layers in order to improve the adaptation of the pre-trained features on the data, but again the accuracy was not better than the reported one, particularly adding a dense layer with 32 units results the accuracy of 0.71.

5 Conclusion

Trying to imitate the way that human beings process new knowledge, the pre-trained language models use model parameters of tasks that have been learned before to initialize the model parameters of new tasks. In this way, the prior knowledge helps new models successfully perform new tasks from old experience instead of from scratch. In this work we solve the task of argument convincingness, meaning that we try to classify the more convincing argument referring the same topic using a pre-trained language model, particularly BERT language model. We conduct our experiments on IBM-EviConv dataset, consisting of pairs of evidence labeled for convincingness, considered to be more challenging than existing alternatives.

Being trained to perform two tasks, masked language model and next sentence prediction, the encoder-based BERT model has been shown to reach state of the art results in argument convincingness classification task. Concretely, we showed that it outperformed the previous models (all models are reported in the original paper Gleize et al., 2019) by reaching the accuracy of 0.75 on the test data.

In this work we used "bert-base-uncased" size variant of BERT, although we believe that there is a room for improvement if the large size variant of BERT is used in case of the availability of computational capacities.

References

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Farra, Noura, Swapna Somasundaran, and Jill Burstein (2015). "Scoring persuasive essays using opinions and their targets". In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 64–74.

Ghosh, Debanjan et al. (2016). "Coarse-grained argumentation features for scoring persuasive essays". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 549–554.

Gleize, Martin et al. (2019). "Are you convinced? Choosing the more convincing evidence with a Siamese network". In: *arXiv preprint arXiv:1907.08971*.

Habernal, Ivan and Iryna Gurevych (2016). "What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation". In: *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1214–1223.

Hidey, Christopher and Kathleen McKeown (2018). "Persuasive influence detection: The role of argument sequencing". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Shnarch, Eyal et al. (2018). "Will it blend? blending weak and strong labeled data in a neural network for argumentation mining". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 599–605.