ACL 2020 START Conference Manager Bill Noble (bnoble)

User Usr

The 58th Annual Meeting of the Association for Computational Linguistics

ACL 2020

Author Response

Title: Tickle me BERT: Using laughter to probe pre-trained encoders for dialogue act recognition
Authors: Bill Noble and Vladislav Maraev

Instructions

The author response period has begun. The reviews for your submission are displayed on this page. If you want to respond to the points raised in the reviews, you may do so in the boxes provided below. *You are not obligated to respond to the reviews.*
Guidelines:

- Remember that reviewing is double-blind.
    - Do not include any information in the response that can identify you or your co-authors.
    - Do not include any URLs in your response.
- Use the response judiciously. If a reviewer has expressed uncertainty about an issue, or is making an incorrect assumption, or has misunderstood a point in the paper, please address these concerns in your author response.
- There is no need to respond to every minor question or suggestion for improvement. In any event, the best strategy is to be polite and professional.
- You are requested **not** to use the response form to debate the reviewers' subjective opinions regarding the merit of your work. Nor should you try to "correct" your paper in any way - either in terms of its basic technical arguments, or in the presentation of those arguments. Above all, the response facility should not be used to report on new results, obtained since the submission deadline closed.
- If the paper is accepted, you will still have ample opportunity to make revisions, i.e., before sending in your camera-ready copy.

Review #1

What is this paper about, what contributions does it make, what are the main strengths and weaknesses?

The paper is mainly about the inclusion of laughter annotation to be used for dialog act recognition using BERT.

The introduction, which misses a section heading, adds laughter as important signal for dialog act recognition. Unfortunately, the authors do not refer to related work as presented in various workshops etc. on paralinguistic signals at venues like Interspeech and ICASSP, where real speech is in the centre. The argumentation with one corpus leaves some doubt about the usefulness of the approach.

The background and the data including processing is described extensively. The model in section 3 is short and it might be hard to reproduce the sequence model. Is the implementation available?

The experiments are well designed and explained. The discussion clearly states that there is “room for further progress”. Since laughter is basically a speech based signal with many functions in different cultures the authors should look into speech corpora and continue their investigation on that basis.

Reasons to accept

The authors present a straightforward approachto integrate laughter in DA recognition.

Reasons to reject

Since paralinguistic signals like laughter are bascially speech signals,using only transcripts seems to be odd

Overall Recommendation: 1.5

Missing References

Please contrastand compare to the workshop series on paralinguistic singals, or books like Björn Schuller, Anton Batliner Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing I am not too deep into this area, bit there might be additional infomation.

Review #2

What is this paper about, what contributions does it make, what are the main strengths and weaknesses?

This paper investigates fine-tuning BERT for dialog act recognition (DAR), the impact of laughter on DAR, and the relationship between the two.

The main strengths are:

1. The experimental results show that BERT learns to represent laughter using a special token during fine-tuning, and it improves the performance of DAR.
2. Impact of pre-training and fine-tuning for BERT-based DAR is measured and compared on the two datasets.

The weaknesses are:

1. This paper provides empirical study on BERT-based DAR, and there is no strong novelty provided by the paper.
2. Even though the title of the paper is about "using laughter", this paper does not emphasize on the point; e.g., in section 4 "Experiments", only subsection 4.1 focuses on using laughter, and the other subsections does not.

Reasons to accept

This paper presents experimental results of BERT trained on two datasets, SWDA and AMI-DA.

Whether laughter information is effective in BERT is verified.

The effectiveness of varying scheme on pre-training and fine-tuning is also verified.

Reasons to reject

This paper presents empirical study on BERT fine-tuned to DAR, and some of the main points (i.e., BERT is better than CNN on DAR, pre-training/fine-tuning for BERT is effective, and etc.) are not very novel because most of them have been verified by other work enough.

Also, what the paper focuses on is not so clear.

The model is compared only with a CNN-based model (not with any other state-of-the-art models).

Overall Recommendation: 2

Questions for the Authors(s)

In Figure 2, what does s^t (maybe, a speaker token) and w_n (maybe, a token of the utt.) mean? It would be better to mention about the notation.

In Table 3, with laughter, the performance of CNN degrades while that of BERT improved. Why would such a result be?

Additional Suggestions for the Author(s)

I think it is better if with new stuffs, the main contribution of this paper is made clearer and more emphasized.

This version provides experimental results of BERT fine-tuned to DAR in various aspects, but the methods or findings are predictable and not pretty novel.

I think something new would be needed for the paper to be accepted to a top conference.

Review #3

What is this paper about, what contributions does it make, what are the main strengths and weaknesses?

The paper describes some experiments into dialogue act tagging using neural networks. It finds that within a RNN sequence model, using BERT as an utterance encoder outperforms using a CNN; that both pre-training and fine-tuning on relevant data are important; and that including laughter as well as lexical material helps performance.

The results on laughter are useful and interesting, and I'm happy to see this kind of work being done in the ACL community - although they seem rather independent of the results on BERT and its training regimes. The main weakness of the paper is that as it stands, it doesn't go very far in investigating the questions about either laughter or pre-training/fine-tuning - lots of avenues could be pursued to clarify the results. The paper feels like a good candidate for a short paper or poster, or for a contribution to a more specific dialogue conference/workshop, but it doesn't seem to contain enough complete research for a full ACL paper.

Reasons to accept

A nice focused contribution on the role of laughter, and the use of pre-training, in dialogue act tagging.

Reasons to reject

With the results and analysis as presented, it's hard for the reader to know what conclusions to draw; waiting for more complete work might be preferable.

Overall Recommendation: 1.5

Questions for the Authors(s)

Given the results on the contribution of laughter, it would be really nice to see some more analysis: which DAs particularly use laughter, how, are there different kinds of laughter, does this fit (or not) with previous linguistic work? Do the CNN and BERT models treat the laughter token in different/sensible ways? Does the attention model in BERT, for example, show that weight is being placed in laughter in a way that seems to fit with the DA patterns of use? Can we say anything about the vectors assigned to laughter tokens (in particular DAs, perhaps?) and what this says about what the models think they "mean" relative to other lexical items?

The results on in/out-of domain pre-training are interesting too, but again some deeper insight would be helpful in order to understand what's really going on. Which DAs are particularly affected, in order to reduce the F-score so much while leaving overall accuracy relatively unaffected? (The analysis of rare words is helpful, but there's more to say here, isn't there?) Are these DAs important in particular ways, particular dialogue settings?

Missing References

The baseline model used for DAR here uses CNNs to encode utterances as vectors, then a RNN to model DA tag sequence; this is clearly explained in Sec 3 but it might be nice to make it clear that this specific approach has been used in other work e.g. Kalchbrenner & Blunsom 2013 (which is cited in the paper but the direct similarity not really pointed out).

Typos, Grammar, and Style

Citing Austin & Urmson as 2009 is perhaps a bit misleading - citing 1955 would be more appropriate!

228 "domain similarly" -> "domain similarity"?

I found Figure 4 in Appendix A very hard to understand - I get the principle of how changes and frequencies are being shown, but this mode of visualisation seems very hard to "get" intuitively. I wonder if separate plots for change and frequency would be easier to process.

It's a tiny thing, but BERT-FT's accuracy on AMI is shown as 66.94 in Table 6 and 66.95 in Table 7

Review #4

What is this paper about, what contributions does it make, what are the main strengths and weaknesses?

This paper experiments with the addition of a laughter token into the vocab for dialogue act classification using BERT. The authors show how the addition of the laughter token can help DA classification on the Switchboard telephone corpus. The authors show how the BERT model responds better than a CNN model to the addition of the token, and for both networks the laughter token helps accuracy. Pre-training is shown not to have any significant effects on performance.

This paper provides a novel approach to an old problem, highlighting the importance of laughter in interaction and dialogue act meaning. The paper is in general very well written and has a good amount of background literature on laughter in human conversation, where the studies show that laughter's function is not always simply due to the expression of being humoured, but has a variety of semantic and pragmatic roles. The analysis of the effect of laughter is thoughtful (though inconclusive) and there is a quantitative error analysis.

The criticisms I have of the paper in its current form is that there is a bit of technical sloppiness in evaluation- it would be good to have a random baseline to see how difficult the task was, and a detailed comparison to the state-of-the-art performances in other papers on DA classification (this may only be possible for the standard 19 dialogues used in Switchboard for DA classification). Also, given the authors have left themselves a spare page, it would have been nice to have seen more examples of where the model is fairing better and worse around laughter events.

Reasons to accept

- A novel integration of laughter into dialogue act classification.
- An interesting overview of laughter in human conservation.

Reasons to reject

- Some sloppiness in the evaluation which may be difficult to correct as there are several key things to collate (see above).
- A lack of qualitative analysis of the model.

Overall Recommendation: 3.5

Submit Response to Reviewers

Use the following boxes to enter your response to the reviews. Please limit the total amount of words in your comments to 900 words (longer responses will not be accepted by the system).

Response to Review #1:

Response to Review #2:

Response to Review #3:

Response to Review #4:

General Response to Reviewers:

Response to Chairs

Use this textbox to contact the chairs directly only when there are serious issues regarding the reviews. Such issues can include reviewers who grossly misunderstood the submission, or have made unfair/unreasonable comparisons or requests in their reviews. Most submissions should not need to use this functionality.

Submit

START Conference Manager (V2.61.0 - Rev. 6009)