From: Daichi Mochihashi <daichi@ism.ac.jp>

Dear x:

As TACL action editor for submission 1304, "Finding Convincing Arguments using Scalable Bayesian Preference Learning", I am happy to tell you that I am accepting your paper subject (conditional) to your making specific revisions within two months.

LIST OF MANDATORY REVISIONS:
- Please address the common issue of the evaluation raised by multiple reviewers.
- Please prepare more explanations and experiments on scalability.
In addition to the main problems from the reviewers listed above, I would like you to add more explanations of Gaussian process preference learning itself, not just relegating it to the citations, to make the paper as self-contained as possible. This is related to our discussion over the decision on this paper: because the strength of this paper lies on the introduction of principled method of GP preference learning that could be also beneficial for other research, it should be explained well for the NLP audience at TACL.

Generally, your revised version will be handled by the same action editor (me) and the same reviewers (if necessary) in making the final decision --- which, *if* all requested revisions are made, will be final acceptance.

You are allowed one to two extra pages of content to accommodate these revisions. To submit your revised version, follow the instructions in the "Revision and Resubmission Policy for TACL Submissions" section of the Author Guidelines at https://transacl.org/ojs/index.php/tacl/about/submissions#authorGuidelines .

Thank you for submitting to TACL, and I look forward to your revised version!

Daichi Mochihashi
The Institute of Statistical Mathematics
daichi@ism.ac.jp

....THE REVIEWS....

Reviewer A:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

4. Understandable by most readers.

ORIGINALITY/INNOVATIVENESS: How original is the approach? Does this paper
break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.
:

3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by properexperiments and are the results of the experiments correctlyinterpreted?:

4. Generally solid work, although there are some aspects of the approach or
evaluation I am not sure about.

MEANINGFUL COMPARISON: Does the author make clear where the presented system
sits with respect to existing literature? Are the references adequate?:

4. Mostly solid bibliography and comparison, but there are a few additional references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of
work), or would it benefit from more ideas or analysis?:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

4. Some of the ideas or results will substantially help other people's ongoing research.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:

5. could easily reproduce the results.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion)
that their software will be available, what is the expected impact of the

software package?:

3. Potentially useful: Someone might find the new software useful for their work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)
that datasets will be released, how valuable will they be to others?:

1. No usable datasets submitted.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Note: after you submit this review form, you'll need to answer a related but different question via a pull-down menu: how long would it take for the authors to revise the submission to be TACL-worthy?
:

4. Worthy: A good paper that is worthy of being published in TACL.

Detailed Comments for the Authors:

The paper proposes a Gaussian process preference learning method to predict argument convincingness.

The method addresses scalability issues of previous methods, making it applicable to larger datasets. The method is shown to outperform previous methods applied for this task. In particular, it achieves significantly better results for ranking, for noisy data, and in an active learning setting.

Technically, it is a solid work, backed by an extensive set of experiments, as well as insightful error analysis. The paper is well written and easy to follow, and there is good coverage of related work.

Two main points for improvement:
(a) In the evaluation section, only the new methods make use of both linguistic and embedding features (ling+Glove), while baseline methods (SVM and BiLSTM) only use one of these sets (ling or Glove). If the authors want to clearly show that the state of the art results were achieved in part by the improved algorithm, their baselines should be the SVM and BiLSTM, each with both ling and Glove features. This is important also from a practical standpoint - SVM and BiLSTM are more accessible methods, so it should be made clear, what is the actual improvement of the new algorithm, over merely adding the Glove features to the SVM, or adding the linguistic features to the BiLSTM. This should be the baseline in any experiment that involves the SVM or the BiLSTM, unless it is shown that adding these features does not

help.

(b) While it is not possible to include complete mathematical description of the proposed method, and the authors do refer to the relevant related work, it would help to provide some more details about the learning method in Section 5. For instance, what q is, and how it is updated.

Minor comments:
Section 3, the paragraph the follows equation (1): which learns the the number of topics than --> rephrase

References: please review your references, make them consistent, fix capitalization etc. Also, fix Reece et al: In proceedings of the ... conference on (*), pages ...

REVIEWER CONFIDENCE:
4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Reviewer B:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:
4. Understandable by most readers.

ORIGINALITY/INNOVATIVENESS: How original is the approach? Does this paper
break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.
:
3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by properexperiments and are the results of the experiments correctlyinterpreted?:
2. Troublesome. There are some ideas worth salvaging here, but the work should really have been done or evaluated differently.

MEANINGFUL COMPARISON: Does the author make clear where the presented

system
sits with respect to existing literature? Are the references adequate?:
2. Only partial awareness and understanding of related work, or a flawed comparison or deficient comparison with other work.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of
work), or would it benefit from more ideas or analysis?:
3. Leaves open one or two natural questions that should have been pursued
within the paper.

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:
3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:
4. could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion)
that their software will be available, what is the expected impact of the software package?:
3. Potentially useful: Someone might find the new software useful for their work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)
that datasets will be released, how valuable will they be to others?:
1. No usable datasets submitted.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Note: after you submit this review form, you'll need to answer a related but different question via a pull-down menu: how long would it take for the authors to revise the submission to be TACL-worthy?
:
2. Leaning against: I'd rather not see it appear in TACL.

Detailed Comments for the Authors:

This paper introduces a scalable Bayesian preference learning method for identifying convincing arguments. The most interesting thing about this method is that it does not require separate methods to produce training data, predict rankings and perform pairwise classifications. When applied to identifying convincing arguments, the method learns from training data composed of pairwise judgments a model for predicting a convincingness score for each argument, from which a ranking of the arguments can be derived. I also find Experiment 1 and the experiment on active learning interesting. Experiment 1 helps the reader understand how the models differ from each other, while the active learning experiment provides suggestive evidence that the model can do a better job than other classifiers for selecting the best instance to annotate.

Overall, the paper is clearly written and is fairly easy to follow. I do think, however, that the major concerns listed below need to be addressed before this paper can be recommended for publication in TACL.

1) Confusing statements

Abstract: "We introduce a scalable Bayesian preference learnng method for identifying convincing arguments in the absence of gold-standard ratings or rankings." This sentence led me to think that this method is unsupervised. Your method does need gold-standard (pairwise) rankings.

In the intro and throughout the paper, you keep emphasizing that the pairwise preferences are noisy. While I understand that the annotations may be noisy because they were obtained by means of crowdsourcing, I don't understand the role played by the noise in the learning process, particularly since there were no results on noise-free data for comparison purposes. In addition, you believe they are noisy, then should we trust the results because they were also evaluated on noisy test data?

2) The major claim lacks substantiation

The main claim is that the proposed preference learning method is scalable, but there isn't any experiment on illustrating how scalable the method is. Section 5 talks about using M inducing points. What value of M did you use? How did you decide that this value should be used? How scalable is this method w.r.t. different values of M? And how sensitive is the method to different values of M?

3) Evaluation issues

I don't understand why the SVM and the BiLSTM were not trained on

"ling+Glove", especially since using both types of features seem to offer better performance. If I understand correctly, the claim is that your model performs better than existing classifiers (LSTM and SVM). The claim is *not* that your model is better than existing classifiers after given a better a feature set. So a fair comparison would involve giving all models the same feature set. In other words, I don't think the comparison in Table 3 is fair at all.

There is no discussion of how the parameters of the SVM and the BiLSTM are set, so it's not clear whether the performance of these classifiers suffered because their parameters were not properly tuned.

It isn't fair to use the SVM/BiLSTM regression results as baselines for the ranking experiments. In fact, it's strange that you needed to cast the SVM/BiLSTM ranking experiments as regression problems and then convert the regression results to ranking results. Many advanced ranker learning algorithms have been developed in the machine learning community in recent years, so why not train them on the PageRank output to directly obtain the ranking?

In Section 6.2, you set the hyperparameters a0 and b0 to certain values "after testing three different settings". Which three settings did you do? Were they obtained by cross validation on the training data? What was the criterion used in selecting these parameters among the three settings?

Overall, it doesn't seem ideal to do ranking experiments on this dataset - as you said, the rankings were automatically derived using PageRank from the pairwise classifications, and in addition, your discussion in Experiment 1 seemed to suggest that PageRank doesn't always produce intuitive rankings. So it seems strange that on one hand, you suggest that the rank produced by your model is better than that of PageRank and on the other hand you use the PageRank rankings as gold standard and evaluate your model's rankings against the PageRank rankings. Since the method (and the resulting model) isn't specifically designed for determining argument convincingness, perhaps it'd be better to evaluate it on tasks where gold-standard rankings are available.

Given these issues, I don't think the paper presented convincing empirical evidence that the proposed model is better than the baselines.

4) Insights into argument convincingness

I am not sure what I learned about the argument convincingness problem from this paper. While I understand that the model seems to produce better results than the baselines, I don't know what linguistic insights I gained into the argument convincingness problem other than the fact that some features were preferred by certain models from Section 6.8. Although the

focus of this paper is the model, without a proper discussion of the linguistic insights that the model provides, the contribution of this paper will merely be an application of an (improved) machine learning technique to a (pairwise) ranking problem with better results.

5) related work

Besides Habernal and Gurevych (2016), there are other papers on automatically determining argument convincingness (e.g.,Liu et al. (ArgMin workshop 2016), Persing & Ng (IJCAI 2017)). Some of them employ different schemes for annotating convincingness than Habernal and Gurevych's and should be discussed. I suggest that the authors did a more thorough investigation of related work.

REVIEWER CONFIDENCE:

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Reviewer C:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

4. Understandable by most readers.

ORIGINALITY/INNOVATIVENESS: How original is the approach? Does this paper
break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
Note that a paper could score high for originality even if the results do not show a convincing benefit.
:

3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by properexperiments and are the results of the experiments correctlyinterpreted?:

4. Generally solid work, although there are some aspects of the approach or
evaluation I am not sure about.

MEANINGFUL COMPARISON: Does the author make clear where the presented

system
sits with respect to existing literature? Are the references adequate?:
4. Mostly solid bibliography and comparison, but there are a few additional references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of
work), or would it benefit from more ideas or analysis?:
4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:
3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:
4. could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion)
that their software will be available, what is the expected impact of the software package?:
1. No usable software released.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)
that datasets will be released, how valuable will they be to others?:
1. No usable datasets submitted.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Note: after you submit this review form, you'll need to answer a related but different question via a pull-down menu: how long would it take for the authors to revise the submission to be TACL-worthy?
:
4. Worthy: A good paper that is worthy of being published in TACL.

Detailed Comments for the Authors:

This paper presents a method for identifying convincing arguments based on Bayesian methods including Preference learning and Gaussian Process classification. The paper uses the publicly available data set developed in (Habernal & Gurevych, ACL 2016), which has two tasks: pairwise classification and argument ranking within a topic. Both tasks are attempted with the adjudicated labels experimented in the original paper and with noisy and possibly conflicting labels. The results show that the current approaches obtain consistent improvements.

The paper is overall well-written and well-structured. The problem is well suited to the techniques proposed. In particular, Gaussian Process Preference Learning is highlighted, a technique which has not been used in NLP, although previous research showed that other variations of Gaussian Processes lead to good results in a series of NLP tasks. The paper applies a stochastic variational inference approach inspired by previous work on Gaussian Processes to handle the size of the data set. Personally, I am surprised that the method is able to scale to use all the linguistic features (>30000), as this is the main limitation of using Gaussian Processes. To this end, I would be interested in a more complete rundown of running times (in addition to the couple of numbers already presented in page 8).

The methods compared against include the two original implementations described in (Habernal & Gurevych, ACL 2016) and the same experimental setup (the results from this paper are better in some cases than the original results, maybe worth explaining why). In addition, a couple of other related methods are compared against: preference learning in the SVM framework as a non-bayesian alternative and Gaussian Process classification as a more conventional non-linear classification alternative. However, I would like to see the comparison being made with SVM's using the same features (ling + Glove) as the new methods. The results show that overall Gaussian Process classification performs best for pairwise classification and GPPL performs best for ranking. The first consideration should be taken into account when making the claims in the abstract and introduction as the classic Gaussian Process performs generally better than the preference learning. I think that, similar to the ULPConvArgStrict/Rank, the tasks in Table 3 should also be split into classification and ranking when presented.

The active learning experiments are a nice addition to the paper. However, it would be preferable to see in Figure 6 the graphs for the entire set of pairs beyond 200.

I find the interpretation section to not be at the level of the rest of the paper. Learning of the ARD lengthscales is one of the key advantages of using Gaussian Processes that facilitates interpretability. I would prefer

to see a more detailed analysis of the top features (lowest lengthscales) in each class of features and any interpretation of why these are useful for predicting convincing arguments. This would also help with the error analysis which is a nice addition to the paper but lacks grounding in the experimental results. Some assumptions are made in that section about convincing/unconvincing arguments (language errors, all caps sentences, multiple exclamation marks, not strongly taking a side) which should have been derived from the data, rather than anecdotical. Also, for this goal Glove embeddings are not useful as they are not interpretable (e.g. what does embedding dimension 19 represent?). For something similar, see the approach of (Preotiuc-Pietro, Lampos, Aletras ACL 2015) where they use word embedding derived topics.

I am not sure of the validity of the claim 'the embeddings are only weakly informative', as table 2 shows that these are highly predictive of the outcome on their own and improve performance when added on top of the other linguistic features. We can also observe that some have very low lengthscales, although the average and median are high for the feature class. I believe this is due to their high coverage when compared to e.g. unigrams - the latter are very sparse, while the embeddings are non-zero for each example.

If space is a concern, I personally did not find the toy data sets to contribute much to the paper + the overview of the paper at the end of the intro can be removed.

I think that the abstract, introduction and conclusion should be toned down and more accurately reflect the findings in the paper:
- in the abstract: "in contrast to previous, work, we avoid the need for separate methods to produce training data, predict rankings and perform pairwise classification" - I do not understand this claim: Training data is obtained from previous work, and previous methods were the same for ranking and classification (whether LSTM or SVM)
- the introduction takes quite long to get to the contributions of this paper. I think some of the initial material which is not novel of this work should be moved to Section 2.
- in the conclusion: "We showed that the method performs well with noisy training data, reducing dependence on a quality control pipeline for crowdsourced data." - I do not find this to be true from the experiments as performance on the second setup is still lower and the patterns of improvement are similar to that of SVMs and LSTMs.

Other considerations:
- I do not think the example in Figure 1 is well chosen as it is quite specific and context is missing (e.g. the ratings for each argument)
- why is the Matern 3/2 kernel used? Were any other kernels tried e.g. the RBF or Matern 5/2?

- what is the motivation for using the inverse scale gamma prior? (Section 5, first paragraph).
- did you try other methods for selecting the inducing points? How about random selection rather than K-means?
- Table 2: move 'UKPConvArgStrict' two rows down.
- ROC AUC different is not measured in percent (Section 6.4), but in absolute 0 to 1 score.
- 'highly statistically significant' - improper use of 'highly', just statistically significant and threshold + the test performed is enough.
- 'rape' is not an emotion term.
- why is 'safer' the most predictive unigram feature, while 'safest' and 'safety' the least?

REVIEWER CONFIDENCE:

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
