Your EMNLP 2018 Submission (Number 1340)

start@sun.softconf.com on behalf of EMNLP 2018 <emnlp2018@softconf.com>

Fri 8/10/2018 2:54 PM

To: Logan Lebanoff <loganlebanoff@Knights.ucf.edu>;

Cc: Logan Lebanoff <loganlebanoff@Knights.ucf.edu>; feiliu@cs.ucf.edu <feiliu@cs.ucf.edu>;

Dear Logan Lebanoff:

On behalf of the EMNLP 2018 Program Committee, we are delighted to inform you that the following submission has been accepted to appear at the conference:

Automatic Detection of Vague Words and Sentences in Privacy Policies

Please be reminded that all accepted papers must be presented at the conference in order to appear in the proceedings. At least one author of each accepted paper must register for EMNLP 2018. (Registration will be open soon on the conference website, emnlp2018.org).

Accepted papers will be presented orally or as a poster. The program committee will decide the format of your presentation and notify you with their decision in early September. There will be no distinction in the proceedings between papers presented orally or as posters.

Papers that are under consideration for another venue must be withdrawn from the other venue if you wish to present them at EMNLP 2018. Alternatively, you may withdraw your paper from EMNLP, which you can do through START.

The Program Committee worked very hard to thoroughly review all the submitted papers. Please repay their efforts, by following their suggestions when you revise your paper.

Your final version needs to be uploaded by August 27, 23:59pm PDT to be included in the conference. The instructions for the final revision will be posted soon on the conference website (emnlp2018.org).

Please upload your final manuscript at the following site:

https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.softconf.com%2Femnlp2018%2Fpapers%2F&data=02%7C01%7Cloganlebanoff%40knights.ucf.edu%7C56d209baeed641a5955708d5fef2b717%7C5b16e18278b3412c919668342689eeb7%7C0%7C0%7C636

695240770280353&sdata=Wo8UH%2FadRdMtW0hLH0Pl%2BrDND2Mv%2BaTnFU2g38BTofU%3D&reserved=0

You will be prompted to login to your START account. If you do not see your submission, you can access it with the following passcode:

1340X-D2P8A5C9D3

Alternatively, you can click on the following URL, which will take you directly to a form to submit your final paper (after logging into your account):

https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.softconf.com%2Femnlp2018%2Fpapers%2Fuser%2Fscmd.cgi%3Fscmd%3DaLogin%26passcode%3D1340X-D2P8A5C9D3&data=02%7C01%7Cloganlebanoff%40knights.ucf.edu%7C56d209baeed641a5955708d5fef2b717%7C5b16e18278b3412c919668342689eeb7%7C0%7C0%7C636695240770280353&sdata=fYJIR1xmzpKtWf9O2GljM%2BTeMLpbcRfDHA2WQlPI8kQ%3D&reserved=0

The reviews and comments are attached below. Again, try to follow their advice when you revise your paper.

Congratulations on your fine work!

Best Regards,

David Chiang, Julia Hockenmaier and Junichi Tsujii
EMNLP 2018
emnlp2018-program@googlegroups.com

==========
EMNLP 2018 Reviews for Submission #1340
==========

Title: Automatic Detection of Vague Words and Sentences in Privacy Policies
Authors: Logan Lebanoff and Fei Liu
==========
REVIEWER #1
==========

What is this paper about, and what contributions does it make?

This paper primarily looks at the problem of automatically detecting vague words and sentences in the context of privacy policies. This paper contributes a new, large dataset of privacy policies annotated for vague words and

phrases, as well as the general vagueness of each sentence. The hypothesis that vague words are context-independent is tested, and the results show that while context-independent representations can represent some aspects of vagueness, context is still important in automatically detecting vagueness. Finally, the paper uses an AC-GAN architecture to predict the vagueness of a sentence, with results surpassing the baseline results.

What strengths does this paper have?

- The paper is very well written, with a good organizational structure and many examples and diagrams along the way that illustrate the ideas being talked about.

- This is the first time that the problem of automatically detecting vagueness has been addressed, and the authors show that an AC-GAN architecture can surpass baselines.

- A large, human-annotated dataset will be released with the paper, allowing others to continue studying the problem of vagueness in privacy policies. This dataset includes annotated vague words and phrases, as well as vagueness sentence scores.

- This paper describes ways to slightly change the AC-GAN architecture to properly handle textual content.

- The figures used are clear and informative.

- Multiple experiments explore the properties of vagueness (such as whether vague words can be detected without context), as well as the best way to automatically detect vagueness.

- This is a very relevant problem with clear societal benefit to be gained from solving it.

- This paper builds well on related work, using previous work to guide the experiments being run.

What weaknesses does this paper have?

- The sentences generated by the AC-GAN (as shown in the supplementary material) are not entirely fluent, and it is unclear why this is; this isn't discussed in the paper. This could be contributing to the result that the Vagueness Only AC-GAN model outperforms the Full AC-GAN model.

Reviewer's Scores

Overall Recommendation: 5

Questions for the Author(s)

Added: Thank you for your additional clarifications in the author response.

Presentation Improvements

- In equation 8, it is unclear what X_{real+fake} is. It would be helpful to define this more precisely in the text.

- The last conclusion under Lessons learned ("Use sentences with simple syntactic structure…") seems unsupported by the experiments described in the paper. Perhaps include more justification for why you are drawing this conclusion (especially when there is little correlation between sentence vagueness and sentence length).

Typos, Grammar, and Style

Line 504: Front quotes improperly formatted in LaTeX

==========
REVIEWER #2
==========

What is this paper about, and what contributions does it make?

This paper introduces a problem of discovering vagueness in text, specifically privacy policies. This is an interesting problem for EMNLP audience and can have potential uses in different areas. Authors prepared a valuable dataset including words and sentences with their vagueness score through amazon mechanical turk platform. They then used LSTM to predict word vagueness and auxiliary-classifier GAN for sentence vagueness prediction task. The main contributions of the paper as states is constructing a corpus for vagueness research and using deep neural networks for predicting vagueness in privacy policies.

What strengths does this paper have?

I believe the dataset that is described in this paper is of interested for many researchers in the field. The problem vagueness, specially in things like privacy policies is an important issue that has been tackled by the authors.

What weaknesses does this paper have?

However, I found the paper interesting, I have a general issue with the framing of the paper. Authors heavily talked about vagueness and privacy policies in the introduction and framed their work in a way that it is solely contributed to that field, however later they spent two pages describing auxiliary-classifier GAN. I wasn't sure if the tweaks in this method was the main contribution. Also, there was no strong justification of using this method over other deep NN

architectures.

Reviewer's Scores

Overall Recommendation: 3

Questions for the Author(s)

I would like to see a few clarifications on the paper: First, did authors add any question to measure participants focus in the process of vagueness scoring? There might be participants who would just click to go forward without paying attention to the task, so it's always good to remove those participants from final dataset. Also it'd be useful if authors could report Kappa's inter-rater reliability for MTurkers.

I think the paper lacks a justification of using A-C GAN for sentence vagueness classification. I am not convinced why authors needed to generate fake sentences given their large dataset. Further, how did authors measure that the fake sentences are actually as vague as the system claims? I think more clarity is needed on convincing readers of using A-C GAN rather than its details.
One final note, would be nice to see whether F-1 scores are significantly different.

Missing References

I would like to see a citation for the first sentence of section 2 (related work) since it is a strong statement.

==========

REVIEWER #3

==========

What is this paper about, and what contributions does it make?

Privacy policies are everywhere online but often difficult to read due to vague terms and phrases that are hard for the average customer to interpret. This paper provides vagueness annotations on a dataset of website privacy policies, at the word and sentence level. The paper then proposes a novel generative adversarial network (GAN) model to classify a given sentence's vagueness level (clear, somewhat clear, vague, extremely vague). At the word-level, the study finds that incorporating context into the word-level classifier improves performance, and also finds that the GAN performs best when classifying clear sentences as compared with vague sentences.

What strengths does this paper have?

---

1. The study provides a clear and well-motivated application of deep learning to a difficult and novel task. I especially appreciated the step-by-step explanation of the AC-GAN model.
2. The error analysis reveals useful linguistic insight, such as the high proportion of nouns tagged as false negatives and the model's difficulty in finding relevant context (Table 5).
3. This kind of analysis is not just useful from a linguistic standpoint - it also sheds light on the bigger social problem of complicated legalese and provides actionable insight for those writing privacy policies. We need more NLP systems like this that can speak truth to power.

What weaknesses does this paper have?

---

1. I found it hard to follow the explanation of the Straight-Through Gumbel generator, particularly as the text refers the reader to Figure 3 for an explanation that seems to be missing.
2. One step of the data filter (Section 3) is to retain only sentences with a vagueness cue word. What kinds of data might be left out by this step? Doing a brief check over the filtered sentences (the ones left out) to make sure that most of these sentences were non-vague would be helpful to know that the filtering wasn't too aggressive.
3. The baselines in Section 6.3 seemed unncessarily complex. Why not use a bag of words logistic regression (similar to Mullenbach et al. 2018)? This would convince the reader that word semantics (not just word presence) are truly important to detecting vagueness.
4. The full GAN model, while technically complex, is shown to underperform. Could the model still be useful for other similar tasks to vagueness detection or would it generally be better to rely on the simpler model (class label only, not discriminating real/fake sentences)?

Reviewer's Scores

Overall Recommendation: 4.5

Questions for the Author(s)

- Can you provide the example sentences used to train annotators? It's important to consider how this might have confounded their labelling performance.
- What kinds of websites were sampled? Section 3 mentions "100 websites" but not the kind of content covered, e.g. if it was biased toward social networks.
- Table 8 shows 0/0 accuracy on extremely vague sentences. This is confusing, because I thought that some of the sentences were tagged as extremely vague. Is that not the case?
- The annotation guidelines in Section 3 don't mention measuring annotator validity, e.g. detecting if any annotator marks every sentence shown with the same score. Did you do any checks to filter such annotators?

AFTER RESPONSE
Not happy with the vague promise to improve the Gumbel estimator - it's distracting from the rest of the paper and makes me wonder what the main contribution is. Also, still not sure what the 0/0 score means - the numerator may be the number of correct responses, but why is the denominator also 0? With the BoW baseline, I worry that there is chance of it doing unusually well and turning the results upside down.

Presentation Improvements

---

1. Show confusion matrix for sentence classification (Table 8) as heatmap.
2. Include reference to ST Gumbel process in Figure 3.

Typos, Grammar, and Style

Several grammar mistakes, such as line 649 "context-agnostic classifier" and line 579 "using using".

--

EMNLP 2018 - https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.softconf.com%2Femnlp2018%2Fpapers&data=02%7C01%7Cloganlebanoff%40knights.ucf.edu%7C56d209baeed641a5955708d5fef2b717%7C5b16e18278b3412c919668342689eeb7%7C0%7C0%7C636695240770280353&sdata=DqrPRypX6JwBGL6KHShmnyol7C%2FpiipXND14hfXSrLc%3D&reserved=0