The 58th Annual Meeting of the Association for Computational Linguistics

ACL 2020

Author Response

Title: ST2: Small-data Text Style Transfer via Multi-task Meta-Learning
Authors: Xiwen Chen and Kenny Zhu

Review #1

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

This paper tackles the problem of text style transfer, focusing specifically on developing a meta-learning framework to allow for style transfer between a large number of different writing styles (as opposed to the usual formality and sentiment tasks). They create (and plan to release) a corpus of literature writing style transfer documents across multiple authors. They evaluate against a set of existing baseline methods with both automatic metrics and human evaluation.

The paper makes a very valid point about the limitation of the text style problem to tasks such as flipping sentiment (which arguably does NOT actually allow you to preserve content...), and moving from formal to informal and vice versa. The authors provide a good amount of detail when describing their baselines and modeling approach, and the fact that they create a new corpus for style transfer from an interesting source is great to see, and could be a good resource for the research community. The analysis of disentangling content and style is also an interesting one.

While the authors argue in their contribution list that their models "outperform the state-of-the-art models in [...] style transfer in terms of content preservation, transfer accuracy, and language fluency", the fact is that the assessment of content preservation is very weak at best. The authors only use BLEU as a heurstic for content preservation, although it is widely agreed to be a very weak measure of content preservation or semantic similarity (see Novikova et al. "Why We Need New Evaluation Metrics for NLG" and Dusek et al. "Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge"). This clear weakness is obvious when inspecting the examples in Table 4: the model outputs from ST^2 clearly NOT preserving the content in most cases. To be fair, this is the case for most of the other outputs as well, but since one of main claims of this work is that the system is able to preserve content as WELL as transfer style, there need to be better measures (they will also be heuristics) for assessing this. Consider trying to identify whether key noun phrases transfer over into the generated texts (see Oraby et al. "Curate and Generate: A Corpus and Method for Joint Control of Semantics and Style in Neural NLG", Reed et al. "Can Neural Generators for Dialogue Learn Sentence Planning and Discourse Structuring?").

It's good to see a human evaluation of the outputs, but this evaluation is also clearly missing any assessment of content preservation. Since evaluating this is hard to do when moving from one surface form to another, human evaluation is an ideal time to try to get some data points for this. Also, there should be more detail given regarding what questions were asked to annotators, what their rubric was, etc.

Finally, while it's interesting that this paper tries to tackle more fine-grained writing styles, and the idea of using literature translations is very interesting, there is really no way to assess whether the writing style is captured in the generated outputs or not, because we have no discrete representation of the input. At the very least, the authors should offer some (even basic) statistics to describe each writing style: for example, vocab size, avg. sentence length, number of adjectives, readability, etc. (see Oraby et al. above) or different LIWC categories. These can be obtained with some off-the-shelf tools, and can then be computed both for the input style AND for the generated outputs for that respective style. This will provide some real intuition into whether the characteristics of the style are being learned and generated.

It's great that the authors release the code, but this paper should also include an appendix of examples.

Reasons to accept

- Good amount of detail in describing baselines and their modeling approach
- New corpus for style transfer (literature domain), code to be released
- Interesting analysis around disentangling content from style

Reasons to reject

- Weak assessment of content preservation, both in the automatic metrics and human evaluation
- Unclear how to really measure if style transfer is happening without any characterization of what the input styles are like
- Need to show many more examples, possibly in an appendix

Overall Recommendation: 2.5

Questions for the Authors(s)

Sec 1:

- Point to the section where you discuss that "most approaches fail on low-resource datasets based on our experiments"

- Throughout the paper, you mention that there are "several styles" and "several style transfer tasks". At least give some examples of what might characterize particular styles: for example, one might be more descriptive (more adjectives), or more verbose (longer sentences), etc. See comments on including some assessment of language styles above.

Sec 3:

- Cite the work you're getting the data from here (around Line 355), and mention the domains
- Where are you getting the literature translations data from? Citation? Link?
- Add a few details about the sentence alignment algorithm you are using (Chen '19)
- How do you pull your 10k sample from the datasets? Is your test set randomly sampled? Need more details here.
- Need more details about this point: "We use the original large dataset to train the language model if the data size is reduced ..."
- Why are you training on the full corpora for GSD and on the training data for LT?
- Human eval: No content preservation? See comments above.
- Human eval: What does this sentence mean? "two evaluators are given the best and worst results generated by all the models to eliminate bias of later-come (?) sentences." Are they expected to calibrate themselves based on how good/bad outputs can be? I imagine the worst outputs will be terrible (just random repeated words like Line 653), and even the best outputs aren't going to be great.
- Not very clear what you mean by "all the baseline models are trained on the single style pair". Elaborate.
- Need to show some statistical significance computations for Table 3.
- Add details here "starting with a well-trained language model, we fine-tune the models for the style transfer task."
- What does this mean: "the way that the model updates its knowledge is parallel rather than sequential..."
- Need to point out fatal flaws with content preservation and an inability to measure it (and offer some heuristic for measuring it). See discussion above.

Missing References

Mentioned a few references in the discussion above.

Typos, Grammar, and Style

The paper needs to be thoroughly proofread. Some typos include:

- state-of-art -> state-of-the-art (at several points in the paper)
- Line 45: broken citation (li2018delete)
- 72: mult-task -> multi-task
- 100: substantially -> substantially
- 222: tc(l) -> ts(l)?
- 224: adversary loss for -> adversarial loss for the
- 279: in to -> into
- 339: either us a semicolon instead of a comma, or add "a" after the comma
- 368: be limited -> limit
- Table 3 caption: larger -> higher

Review #2

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

The paper proposes a multi-task meta-learning method for text style transfer task. Specifically, the proposed method adapts meta-learning framework to existing models (i.e. CrossAlign model and variational autoencoder model), to alleviate the problem of lack of parallel training data.

Strengths: (1) This paper is the first to apply meta-learning method on text style transfer task.. (2) According to the experiments evaluation, the method performs best among all baselines, and a novel writing style transfer dataset will be released.

Weakness: (1) This paper incrementally apply existing models to meta-learning framework for small-data text style transfer task. (2) Wide coverage of this paper introduces the existing models in Section(2). The focus should be on the proposed model.

Reasons to accept

(1) This paper applies meta-learning framework on text style transfer task, which is novel and reasonable. (2) This paper conducts sufficient experiment, and the proposed method performs best among start-of-the-art models.

Reasons to reject

(1) The incremental method of this paper fails to further improve the performance based on the characteristic of meta-learning framework and text style transfer task. (2) The description of proposed method is not detail.

Overall Recommendation: 2.5

Typos, Grammar, and Style

In line 97, this is the first work that…. In line 107, which is the first of its kind….

Review #3

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

Summary and Contributions. This paper proposes to go beyond the binary positive/negative and formal/informal style transfer work, by incorporating other aspects of style, some with a binary structure (e.g., standard vs simple style), and others with a less structured author style (e.g., two individual writers). A novel dataset with various styles is collected, and apply a Meta-Agnostic Meta-Learning (MAML) framework to propose a Multi-Task Text Style Transfer (ST^2) algorithm. The paper conducts automated evaluation for BLEU, perplexity, transfer accuracy, and a human fluency ranking.

Strengths. This paper proposes a natural extension of the style transfer task to look beyond simple binary positive/negative and formal/informal transfer. To experiment with this, a novel dataset is collected of two novel resources: multiple translations of a piece of literature, and a grouped subset of popular, smaller datasets that on their own, are not sufficient in size for previous approaches to style transfer. The proposed multi-task algorithm outperforms most other models on the new datasets.

Weaknesses. Some of the evaluation metrics are lower than or on par with the baseline or previous models, e.g., BLEU, PPL, Human (worse or tied) for the LT and GSD datasets in Table 3, and BLEU and PPL (tied) in Table 5. The human evaluation on fluency could be more detailed (separate out fluency from naturalness, introduce other considerations).

Reasons to accept

This innovative dataset opens the door to expansion of the style transfer task by gathering all sorts of varied texts. The work in this paper will keep forward-momentum in the field, rather than making tiny increments to the existing problems with the existing resources. The new resources, model, and extension

of the style transfer task would be a great benefit to the NLG community at large.

Reasons to reject

Despite my remarks about the automated evaluation metrics, this paper's strengths greatly outweigh the weaknesses. In fact the authors do a reasonable job of addressing these shortcomings themselves, stating that the better-scoring models utilize templates, but nonetheless warrants further evaluation, which they conduct to examine the effects of pretraining.

Overall Recommendation: 4

Typos, Grammar, and Style

In the second paragraph of the intro, 'li2018delete' is likely missing a \cite{} tag.

Review #4

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

The problem that they are addressing is generating text style transfer while having smaller amounts of training data for each style. They do this by using a meta-learning framework applied to existing state-of-the-art models. They have good descriptions of the models that they use for their framework and they used multiple datasets to show that their model works on multiple types of style-transfer tasks. Some weaknesses are that they assume that BLEU is a good measurement of content preservation, but this is not necessarily so. They also present an automatic evaluation of transfer accuracy without demonstrating that it is effective for their given task. Especially for their writing style transfer, which is very different than what the classifier was tested within the cited work.

Reasons to accept

They are using a framework that hasn’t been applied to this problem and they show that it improves existing results. Their human annotation results are quite good, as are their other metrics. It is an interesting step forward for a problem that is definitely relevant and important.

Reasons to reject

I’m not confident about their evaluation metrics. BLEU is not the best metric for determining content preservation, especially when calculated using the original sentences which can be quite different even if the content is preserved. They don’t show that the writing style between the different authors is clearly identifiable or that ACC is capable of distinguishing between them. Having the human annotators also determine if the content is preserved and that the style is properly transferred would probably help this.

Overall Recommendation: 3.5

Questions for the Authors(s)

I would be interested in seeing some examples from the output that shows it working well. I appreciate being given random examples, but I'm also curious about what the model generates in the best-case scenario.

Typos, Grammar, and Style

Line 045: citation didn't get properly generated. Line 072: multi-task. Line 097: "...this is the first work…". Line 096 "state-of-the-art".