Diversified evaluation of text simplification through extrinsic tasks
Tianchi Zuo (The University of Tokyo), Naoki Yoshinaga (IIS, The University of Tokyo)

Summary
We propose to evaluate text simplification through the extrinsic tasks:
- use output scores of task models trained with data simplified by the target model
- We evaluate correlation between scores and human judgement

How to evaluate text simplification?
Goal: Simplify text while keeping the meaning.

Input: Avatar is maddening
Output1: Avatar is bad
Output2: Avatar is very bad

Human judgement:
- Meaning
- Grammar
- Simplicity

Not reproducible
Automatic evaluation
- BLEU [Papeneli+ 02], SARI [Xu+ 16]

Ignore the importance of individual words/phrases

Evaluate text simplification via tasks
Idea: Examine models for various tasks trained from data simplified by the target model.
- Simplified sentences can be easier to process by computers as well as human.
- Outputs of simplified model indicate the effectiveness of text simplification.

Sentiment classification as the extrinsic task
Step 1. Simplify task datasets w/ target simplification model.
Step 2. Train a model from simplified datasets.
Step 3. Obtain scores for each example in simplified dataset.

Training/test datasets
Simplified training/test datasets
Simplified model

“Avatar is bad” Negative (0.7) < “Avatar is very bad” Negative (0.8)

We can expect better simplification increases the probability of gold outputs.

Experiments
Evaluate our method in terms of correlation between model outputs and human judgement.

Extrinsic Tasks for evaluation
- Sentiment analysis (SST-2)
- Natural language inference (MNLI train+SNLI test)
- Language modeling (Simple-wikipedia)

Text simplification model
- ACCESS [Martin+ 19]

Meaning | Grammar | Simplicity
------- | ------- | -------
Senti. Analysis | -0.04 | 0.16 | 0.03
Natural language inference | 0.05 | -0.10 | 0.01
Lang. modeling | -0.16 | 0.10 | 0.11

Conclusion
- The results of language modeling shows that this task can slightly reflect text simplicity.
- Other tasks seems not so useful as expected.

Future work
1. Try other simplification models for comparison.
2. Leverage the gain of performance of simplified model as metric.
3. Try other models for each extrinsic tasks.
4. Consider task-oriented text simplification.