Does It Capture STEL? A Modular, Similarity-Based Linguistic Style Evaluation Framework

Anna Wegmann and Dong Nguyen

Department of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands

NLP and Society Lab

Why?

Style is an integral part of language:
- influences perception
- relevant for NLU & NLG

Previously, in style evaluation:
- Authorship Attribution

↯ **content ∞ style**

e.g., [Zangerle et al., Overview of the Style Change Detection Task at PAN 2020, 2020]

- Style Classification

↯ **task-specific**

e.g., [Kang and Hovy, Style is NOT a Single Variable: Case Studies for Cross-Style Language Understanding, 2021]

- Overall

↯ **multiple style definitions**

e.g., [Biber and Conrad, Register, Genre and Style, 2009; Crystal and Davy, Investigating English Style, 1969]

How?

formal	informal
Are you one of their fans?	**r u** a fan of them or something?
Oh, and also that young physician got an unflattering haircut	Oh **yea** and that young **dr** got a bad haircut

content-controlled

r u a fan of them or something? | Are you one of their fans?

Oh, and also that young physician got an unflattering haircut | Oh **yea** and that young **dr** got a bad haircut

STEL is modular:

- formal/informal
  Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset , 2018]
- (no) numb3r substitution
  based on **Reddit** utterances
- complex/simple
  Wikipedia paraphrases from [Xu et al., Optimizing Statistical Machine Translation for Text Simplification, 2016]
- (no) con'raction
  based on **Wikipedia** abstracts

Selected Results:

	all	formal	simple	nb3r	c'tion
BERT uncased	0.74	0.79	0.65	0.90	0.90
BERT cased	0.77	0.82	0.68	0.92	1.00
RoBERTa	0.61	0.63	0.54	0.62	0.98
BERT uncased NSP	0.66	0.72	0.59	0.67	0.70
BERT cased NSP	0.71	0.79	0.60	0.77	0.96
share cased	0.56	0.55	0.53	0.50	1.00
word length	0.58	0.53	0.59	0.50	0.94

Table: Accuracy of style methods. Random performance corresponds to 0.5.

What?

Towards a general style evaluation benchmark, we contribute
- **the modular, content-controlled, similarity-based STEL framework**
- **1830 task instances across 4 dimensions**
- **Baseline results for 18 models**

Next?

- additional STEL dimensions
- testing for 'style over content'

Who?

Code: `github.com/nlpsoc/STEL`
Email: `a.m.wegmann@uu.nl`, `d.p.nguyen@uu.nl`

References