Does It Capture STEL?

A Modular, Similarity-based Linguistic Style Evaluation Framework

Anna Wegmann and Dong Nguyen
Utrecht University, Utrecht, the Netherlands

November 2021, **EMNLP**, Online & Dominican Republic

Utrecht University

NLP and Society Lab

Style in Natural Language Processing

Style is an integral
part of language:

Style in Natural Language Processing

Style is an integral
part of language:

- influences perception

e.g., [El Baff et al., Analyzing the Effect of Style in News Editorial Argumentation, 2020]

Style in Natural Language Processing

Style is an integral
part of language:

- influences perception
  e.g., [El Baff et al., Analyzing the Effect of Style in News Editorial Argumentation, 2020]

- relevant for NLU & NLG
  e.g., [Danescu-Niculescu-Mizil et al., Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs, 2011], [Ficler and Goldberg, Controlling Linguistic Style Aspects in Neural Language Generation, 2017]

Style in Natural Language Processing

Style is an integral part of language:

- influences perception

  e.g., [El Baff et al., Analyzing the Effect of Style in News Editorial Argumentation, 2020]

- relevant for NLU & NLG

  e.g., [Danescu-Niculescu-Mizil et al., Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs, 2011], [Ficler and Goldberg, Controlling Linguistic Style Aspects in Neural Language Generation, 2017]

General Evaluation of Style Measures?

Style in Natural Language Processing

Style is an integral part of language:

- influences perception
  e.g., [El Baff et al., Analyzing the Effect of Style in News Editorial Argumentation, 2020]
- relevant for NLU & NLG
  e.g., [Danescu-Niculescu-Mizil et al., Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs, 2011], [Ficler and Goldberg, Controlling Linguistic Style Aspects in Neural Language Generation, 2017]

General Evaluation of Style Measures?

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

↯ content ∞ style

Style in Natural Language Processing

Style is an integral part of language:

- influences perception
  e.g., [El Baff et al., Analyzing the Effect of Style in News Editorial Argumentation, 2020]
- relevant for NLU & NLG
  e.g., [Danescu-Niculescu-Mizil et al., Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs, 2011], [Ficler and Goldberg, Controlling Linguistic Style Aspects in Neural Language Generation, 2017]

General Evaluation of Style Measures?

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

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

  ↯ task-specific

Style Measurement Evaluation

Difficulties:

- ↯ content ∞ style

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

- ↯ task-specific

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

Style Measurement Evaluation

Difficulties:

- ↯ content ∞ style

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

- ↯ task-specific

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

- ↯ multiple style definitions

  e.g., [Biber and Conrad, Register, Genre and Style, 2009; Labov, The Social Stratification of English in New York City, 2006; Crystal and Davy, Investigating English Style, 1969; Xu, From Shakespeare to Twitter: What are Language Styles all about?, 2017]

Style Measurement Evaluation

Difficulties:

- ↯ content ∞ style

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

- ↯ task-specific

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

- ↯ multiple style definitions

  e.g., [Biber and Conrad, Register, Genre and Style, 2009; Labov, The Social Stratification of English in New York City, 2006; Crystal and Davy, Investigating English Style, 1969; Xu, From Shakespeare to Twitter: What are Language Styles all about?, 2017]

Our Proposal:

⇒ STyle EvaLuation Framework (STEL)

Style Measurement Evaluation

Difficulties:

- content ∞ style
  e.g., [Zangerle et al., Overview of the Style Change Detection Task at PAN 2020, 2020]
- task-specific
  e.g., [Kang and Hovy, Style is NOT a Single Variable: Case Studies for Cross-Stylistic Language Understanding, 2021]
- multiple style definitions
  e.g., [Biber and Conrad, Register, Genre and Style, 2009; Labov, The Social Stratification of English in New York City, 2006; Crystal and Davy, Investigating English Style, 1969; Xu, From Shakespeare to Twitter: What are Language Styles all about?, 2017]

Our Proposal:

- content-controlled

⇒ STyle EvaLuation Framework (STEL)

Style Measurement Evaluation

Difficulties:

- content ∞ style
  e.g., [Zangerle et al., Overview of the Style Change Detection Task at PAN 2020, 2020]
- task-specific
  e.g., [Kang and Hovy, Style is NOT a Single Variable: Case Studies for Cross-Stylistic Language Understanding, 2021]
- multiple style definitions
  e.g., [Biber and Conrad, Register, Genre and Style, 2009; Labov, The Social Stratification of English in New York City, 2006; Crystal and Davy, Investigating English Style, 1969; Xu, From Shakespeare to Twitter: What are Language Styles all about?, 2017]

Our Proposal:

- content-controlled
- similarity-based

⇒ STyle EvaLuation Framework (STEL)

Style Measurement Evaluation

Difficulties:

- ↯ content ∞ style
  e.g., [Zangerle et al., Overview of the Style Change Detection Task at PAN 2020, 2020]
- ↯ task-specific
  e.g., [Kang and Hovy, Style is NOT a Single Variable: Case Studies for Cross-Stylistic Language Understanding, 2021]
- ↯ multiple style definitions
  e.g., [Biber and Conrad, Register, Genre and Style, 2009; Labov, The Social Stratification of English in New York City, 2006; Crystal and Davy, Investigating English Style, 1969; Xu, From Shakespeare to Twitter: What are Language Styles all about?, 2017]

Our Proposal:

- content-controlled
- similarity-based
- modular

⇒ STyle EvaLuation Framework (STEL)

STEL Task

	1	2	
Anchor (A)	r u a fan of them or something?	Are you one of their fans?	Do S and A "match"?
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut	

STEL Task

	1	2
Anchor (A)	r u a fan of them or something?	Are you one of their fans?
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut

STEL Task

	1	2	
Anchor (A)	r u a fan of them or something?	Are you one of their fans?	paraphrase
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut	

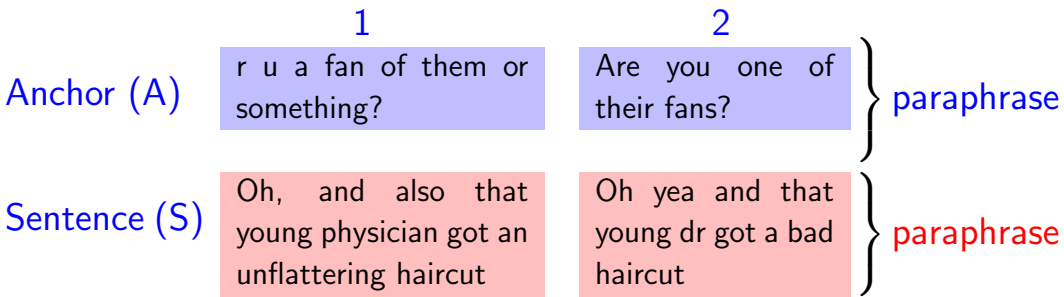
STEL Task

	1	2	
Anchor (A)	r u a fan of them or something?	Are you one of their fans?	paraphrase
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut	

STEL Task

	1	2	
Anchor (A)	r u a fan of them or something?	Are you one of their fans?	paraphrase
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut	paraphrase

STEL Task is content-controlled



STEL Task

	1	2	
Anchor (A)	r u a fan of them or something?	Are you one of their fans?	Do S and A "match"?
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut	

STEL Task

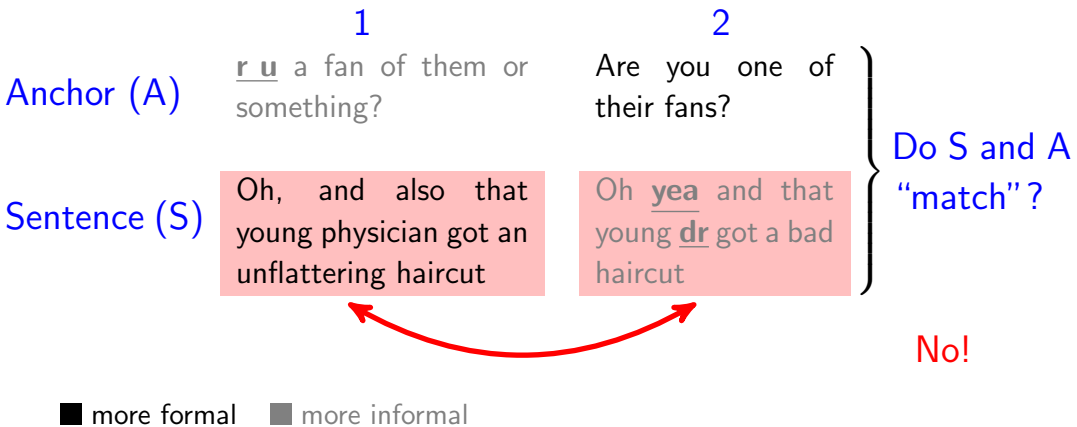
	1	2	
Anchor (A)	**r u** a fan of morethem or something?	Are you one of their fans?	Do S and A "match"?
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh **yea** and that young **dr** got a bad haircut	

STEL Task

	1	2	
Anchor (A)	**r u** a fan of them or something?	Are you one of their fans?	Do S and A "match"?
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh **yea** and that young **dr** got a bad haircut	

■ more formal ■ more informal

STEL Task


1
2
Anchor (A)
r u a fan of them or something?
Are you one of their fans?
Sentence (S)
Oh, and also that young physician got an unflattering haircut
Oh yea and that young dr got a bad haircut
Do S and A "match"?
No!
more formal
more informal

STEL Task Generation*

*Examples shortened

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

- **formal/informal**

Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 2018]

STEL Task Generation*

*Examples shortened

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

r u a fan of them or something?

Oh, and also that young physician got an unflattering haircut

Are you one of their fans?

Oh **yea** and that young **dr** got a bad haircut

- **formal/informal**

Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 2018]

STEL Task Generation*

*Examples shortened

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

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

- **formal/informal**

Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 2018]

STEL Task Generation*

*Examples shortened

simple	complex
These rock formations are made of sandstone.	These rock formations are **composed** of sandstone.
The Odyssey is an old poem written by Homer.	The Odyssey is an **ancient** poem **attributed** to Homer.

These rock formations are **composed** of sandstone.

These rock formations are made of sandstone.

The Odyssey is an **ancient** poem attributed to Homer.

The Odyssey is an old poem written by Homer.

- formal/informal

Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 2018]

- **complex/simple**

Wikipedia paraphrases from [Xu et al., Optimizing Statistical Machine Translation for Text Simplification, 2016]

STEL Task Generation*

*Examples shortened

no nb3r	nb3r
<3 friends forever	<3 friends **4**ever
Dude 30$ is heaps cheap	**D00d** 30$ is heaps cheap

<3 friends forever

<3 friends **4**ever

D00d 30$ is heaps cheap

Dude 30$ is heaps cheap

- formal/informal

  Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 2018]

- **(no) numb3r substitution**

  based on **Reddit** utterances

- complex/simple

  Wikipedia paraphrases from [Xu et al., Optimizing Statistical Machine Translation for Text Simplification, 2016]

STEL Task Generation*

*Examples shortened

no con'raction	con'raction
It has become one of the world's most significant cities.	**It's** become one of the world's most significant cities.
Will does not refer to any particular desire.	Will **doesn't** refer to any particular desire.

It has become one of the world's most significant cities.

It's become one of the world's most significant cities.

Will does not refer to any particular desire.

Will **doesn't** refer to any particular desire.

- formal/informal

  Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 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

STEL Task Generation*

*Examples shortened

Style 1	Style 2
Anchor Sentence 1	Anchor Sentence 2
Alternative Sentence 2	Alternative Sentence 1

Paraphrases

Anchor Sentence 1 Anchor Sentence 2

Alternative Sentence 1 Alternative Sentence 2 ✓

- formal/informal

  Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 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

STEL Task Generation*

*Examples shortened

Style 1	**Style 2**
Anchor Sentence 1	Anchor Sentence 2
Alternative Sentence 2	Alternative Sentence 1

Paraphrases

Anchor Sentence 1 Anchor Sentence 2

Alternative Sentence 1 Alternative Sentence 2

STEL is modular:

- formal/informal

  Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 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

STEL Task Generation*

*Examples shortened

Style 1	Style 2
Anchor Sentence 1	Anchor Sentence 2
Alternative Sentence 2	Alternative Sentence 1

Paraphrases

Anchor Sentence 1 Anchor Sentence 2

Alternative Sentence 1 Alternative Sentence 2 ✓

STEL is modular:

complex
- formal/informal
  Yahoo! Answers paraphrases from [Rao and Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, 2018]
- complex/simple
  Wikipedia paraphrases from [Xu et al., Optimizing Statistical Machine Translation for Text Simplification, 2016]

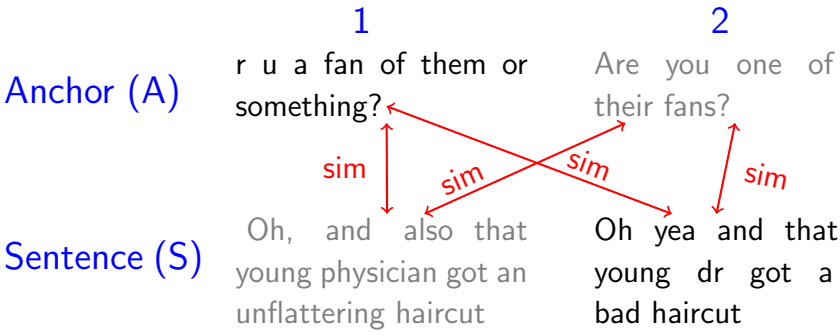
simple
- (no) numb3r substitution
  based on **Reddit** utterances
- (no) con'raction
  based on **Wikipedia** abstracts

Model Evaluation

	1	2
Anchor (A)	r u a fan of them or something?	Are you one of their fans?
Sentence (S)	Oh, and also that young physician got an unflattering haircut	Oh yea and that young dr got a bad haircut

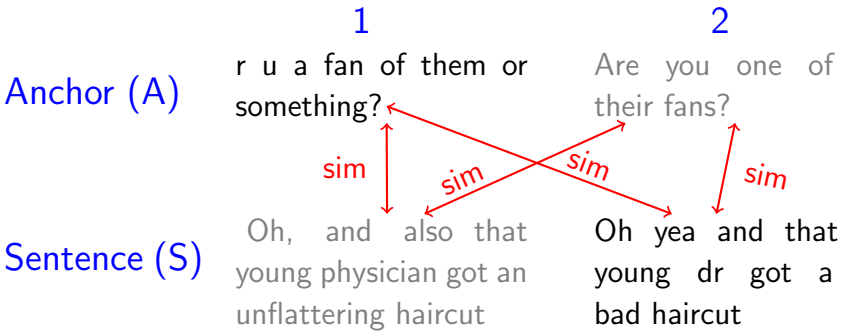
How should models decide?

Model Evaluation



How should models decide?

Model Evaluation is Similarity-Based



$\longrightarrow$ Any method that compares two sentences can be evaluated

Selected Results

- BERT base models perform the best

	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	1.0

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

For 14 more models & methods, see our paper ...

Selected Results

- BERT base models perform the best

- Casing encodes style info

	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	1.0
share cased	0.56	0.55	0.53	0.50	1.0

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

For 14 more models & methods, see our paper ...

Selected Results

- BERT base models perform the best

- Casing encodes style info

- word length predicts complexity similar to [Paetzold and Specia, SemEval 2016 Task 11: Complex Word Identification, 2016]

	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	1.0
share cased	0.56	0.55	0.53	0.50	1.0
word length	0.58	0.53	0.59	0.50	0.94

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

For 14 more models & methods, see our paper ...

Summary

- modular, content-controlled similarity-based STEL framework
- 1830 tasks across 4 dimensions
- Baseline results for 18 models
  $\rightarrow$ cased BERT encodes style information

Summary

- modular, content-controlled similarity-based STEL framework
- 1830 tasks across 4 dimensions
- Baseline results for 18 models
  $\rightarrow$ cased BERT encodes style information

Future Work

- additional STEL dimensions
- tasks testing for "style over content"

Does It Capture STEL? @ EMNLP 2021

- modular, content-controlled similarity-based STEL framework
- 1830 tasks across 4 dimensions
- Baseline results for 18 models
  $\rightarrow$ cased BERT encodes style information
- additional STEL dimensions
- tasks testing for "style over content"

Data & Code: `github.com/nlpsoc/STEL`

`@anna_wegmann` `@dongng`