The Daunting Task of Real-World Textual Style Transfer Auto-Evaluation

Richard Yuanzhe Pang
New York University
Work done at the University of Chicago and TTIC

arXiv: 1910.03747

Task

X_0, X_1: Two non-parallel corpora of different "styles"
$\mathbf{x}_t^{(i)}$: ith sentence of style t

<u>Want</u> $\widetilde{\mathbf{x}}_0^{(i)}$: sentence with style 1 but the content of $\mathbf{x}_0^{(i)}$ $\widetilde{\mathbf{x}}_1^{(i)}$: sentence with style 0 but the content of $\mathbf{x}_1^{(i)}$

Lack of parallel corpora => Need unsup learning criteria and auto-evaluation metrics

Background: "Supervised" Eval Based on Human-Written "Gold-Standards"

Model	BLEU	Acc
Shen et al. (2017)		
CAE	4.9	0.818
CAE	6.8	0.765
Fu et al. (2018)		
Multi-decoder	7.6	0.792
Style embed.	15.4	0.095
Li et al. (2018)		
Template	18.0	0.867
Delete/Retrieve	12.6	0.909

Model	BLEU	Acc
Yang et al. (2018)		
LM	13.4	0.854
LM + classifier	22.3	0.900
Pang and Gimpel (2018)		
CAE + losses (M6)	22.5	0.843
CAE + losses (M6)	16.3	0.897
Untransferred	31.4	0.024

<u>**BLEU**</u> is between 1000 Yelp transferred sentences and human written gold-standard references (Li et al., 2018)

<u>**Acc**</u> Post-transfer style classification accuracy (computed by pretrained classifier)

<u>**Observation**</u>
(1) BLEU has inverse relationship with Acc
***(2) Untransferred* sentences have highest BLEU**

Unreliable and costly

Background: Existing Auto-Evaluation Metrics

	Pang and Gimpel (2018)	Mir et al. (2019)
1. Acc (post-transfer accuracy)	How often was a pretrained style-classifier convinced of transfer?	
2. Sim (semantic similarity)	• (i) Embed sentences by avg word embeddings (GloVe, 300d) weighted by idf • (ii) Sim is the avg of the cos sim over all original/transferred sentence pairs	• (i) Remove style words from original sentence and transferred sentence using a style lexicon (by classifier), and then replace those words with <customstyle> labels • (ii) Use METEOR and Earth's Mover's Distance to compute Sim
3. PP (fluency or naturalness) *Perplexity is distinct from fluency, but correlated*	Measured by perplexity (by language model trained on <u>concatenation of two corpora</u>)	Measured by perplexity (by language model trained on <u>target corpora</u>)

Problem 1 (of recent research): Style transfer TASKS

Recent research focuses on <u>*operational* transfer</u> like Yelp sentiment transfer (vocabs of two styles are similar; can use simple classifier to determine style); DOES NOT represent REAL-WORLD style transfer!

REAL-WORLD applications	Examples
1. Writing assistance	Formality transfer; politeness transfer; dialogue
2. Author obfuscation and anonymity	… so that authors can stay relatively anonymous in heated political discussions
3. For artistic purposes	Transfer modern article to old literature styles
4. Adjusting reading difficulty in education	Generating passages of same content, but of different difficulty levels appropriate to different age groups
5. Data augmentation to fix dataset bias	In sentiment classification, "romantic"=>positive, "horror"=>negative; can generate sentences with flipped sentiment BUT same content; Can also apply to social bias issues (gender, race, nationality, etc.)

Style transfer task	**CONTENT-related words**	**STYLE-related words**
#5 on the left: data augmentation (by sentiment transfer) to fix movie review dataset bias	Positive: "romantic" Negative: "horror"	Positive: "amazing" Negative: "awful"
#3 on the left: Dickens <-> Modern literature transfer	Dickens: "English farm" "horses" Modern: "vampire" "pop music"	Dickens: "devil-may-care" "flummox" Modern: "chill"
	SHOULD BE **LEFT UNCHANGED**	SHOULD **CHANGE**

Different styles' original corpora have different vocabs
=> Hard to distinguish content-related words from style-related words

<u>**But**</u> current research focuses on Yelp sentiment transfer (vocab of two styles are similar); DOES NOT represent REAL-WORLD style transfer!

Problem 2 (of recent research): Metrics

<u>Dickens style → Modern style</u>
Original sentence: ***Oliver deemed the gathering in York a great success.***
Real-world style transfer: ***Oliver thought the gathering was successful.***
Operational style transfer (recent research): ***Karl enjoyed the party in LA.***

Corpus-specific content proper nouns	"Oliver", "York": Should stay!
Other corpus-specific content words	"English farm", "horses": Should stay!
Style words	"deemed", "gathering": Should change!

Sim	• **Problem**: Should not include content words in computing Sim • Option 0 (incorrect): Use classifier to determine style lexicon, and mask out style keywords • Option 1: Manually create a list of style lexicon, and mask out style keywords • Option 2: Keep the words as they are, and compute Sim directly
Acc	• **Problem**: Should not include content words in classifier
PP	• **Problem**: Should not include content words in computing PP • **Another problem**: Very low PP does not indicate fluency, need to punish very low PP

Problem 3: Tradeoff and Aggregation of Scores

Pang and Gimpel (2018): Negative relationship b/w Sim and Acc; Mostly positive relationship b/w PP and Sim => TRADEOFF

A = Acc, B = Sim, C = PP
Score = f(A,B,C) for ease of model selection and comparison; Can train f with human annotations of pairwise comparison

<u>Bibliography</u>
Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.
Yuanzhe Pang and Kevin Gimpel. 2018. Learning criteria and evaluation metrics for textual transfer between non-parallel corpora. *arXiv preprint arXiv:1810.11878.*
Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *In Advances in Neural Information Processing Systems 30*, pages 6833–6844. Curran Associates, Inc.