Reformulating Unsupervised Style Transfer as Paraphrase Generation

Kalpesh Krishna, John Wieting, Mohit Iyyer

Some slides from EMNLP20 presentation: http://style.cs.umass.edu/

Issue with recent work in style transfer

Models significantly modify content of sentences

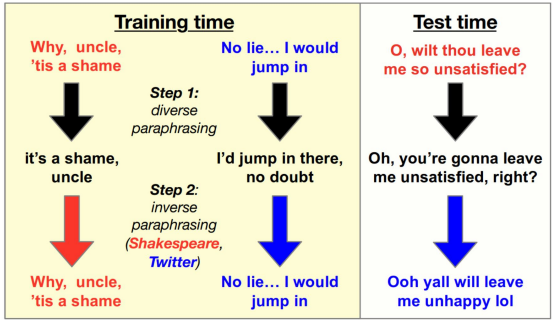
Relaxed ↔ Annoyed	
Relaxed	Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night 💗🎄✨
Annoyed	Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend 😡😡😡
Annoyed	Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month 😒
Relaxed	Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month 😊
Male ↔ Female	
Male	Gotta say that beard makes you look like a Viking...
Female	Gotta say that hair makes you look like a Mermaid...
Female	Awww he's so gorgeous 😍 can't wait for a cuddle. Well done 😘 xxx
Male	Bro he's so f***ing dope can't wait for a cuddle. Well done bro
Age 18-24 ↔ 65+	
18-24	You cheated on me but now I know nothing about loyalty 😂 ok
65+	You cheated on America but now I know nothing about patriotism. So ok.
65+	Ah! Sweet photo of the sisters. So happy to see them together today .
18-24	Ah 😂 Thankyou 💞 #sisters ❤️ happy to see them together today

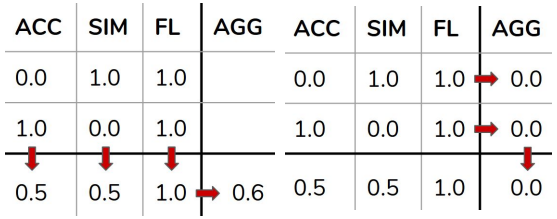
Outputs from Lample et al. 2019, ICLR

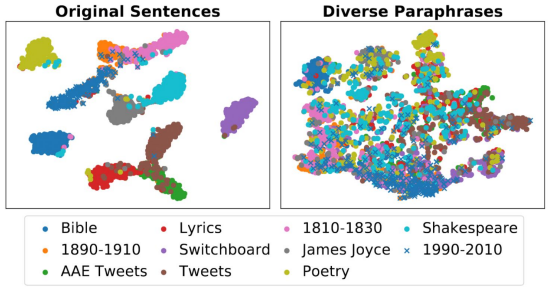
Simple new **state-of-the-art unsupervised** algorithm, models style transfer as controlled paraphrase generation

23-paper survey of evaluation methods, **improvements**

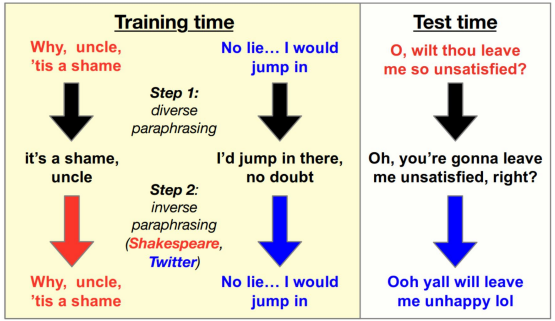
New corpus of **15M** sentences, **11** diverse styles (Tweets, Bible, Poetry, speech transcripts etc.)



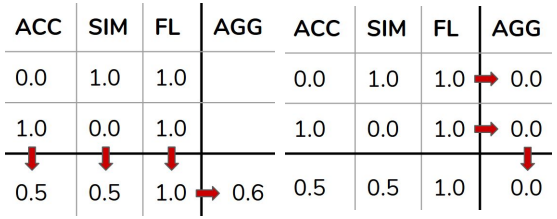




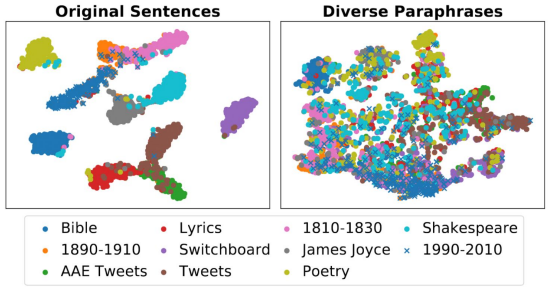
Simple new **state-of-the-art unsupervised** algorithm, models style transfer as controlled paraphrase generation



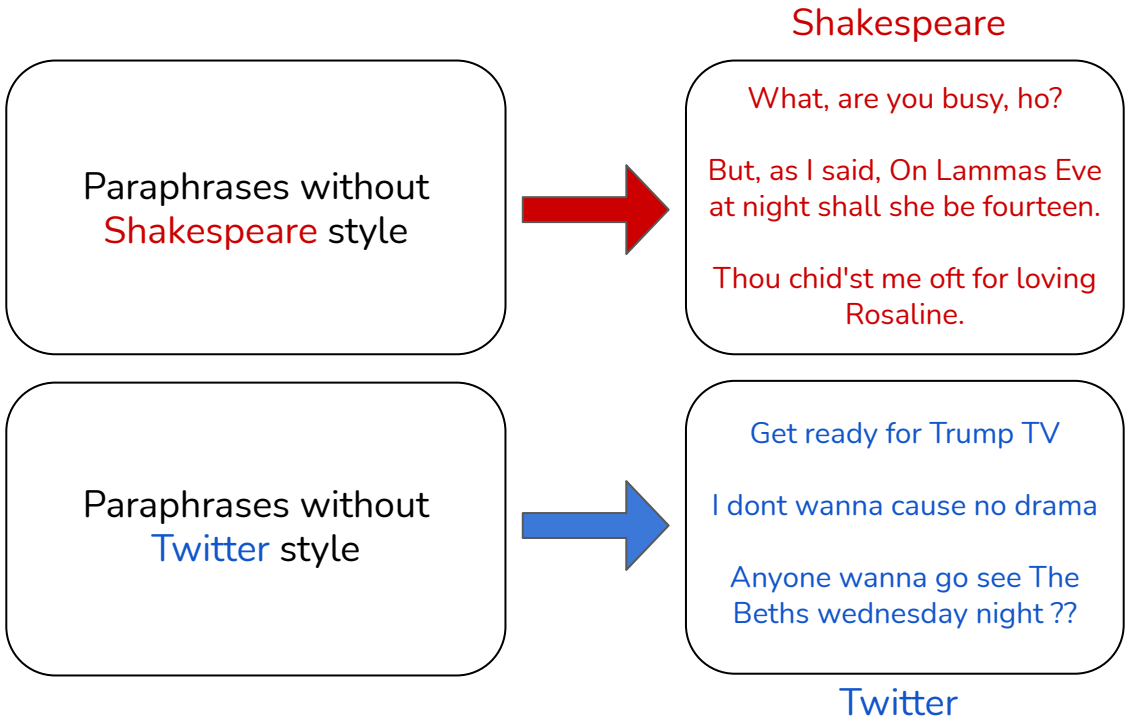
23-paper survey of evaluation methods, **improvements**



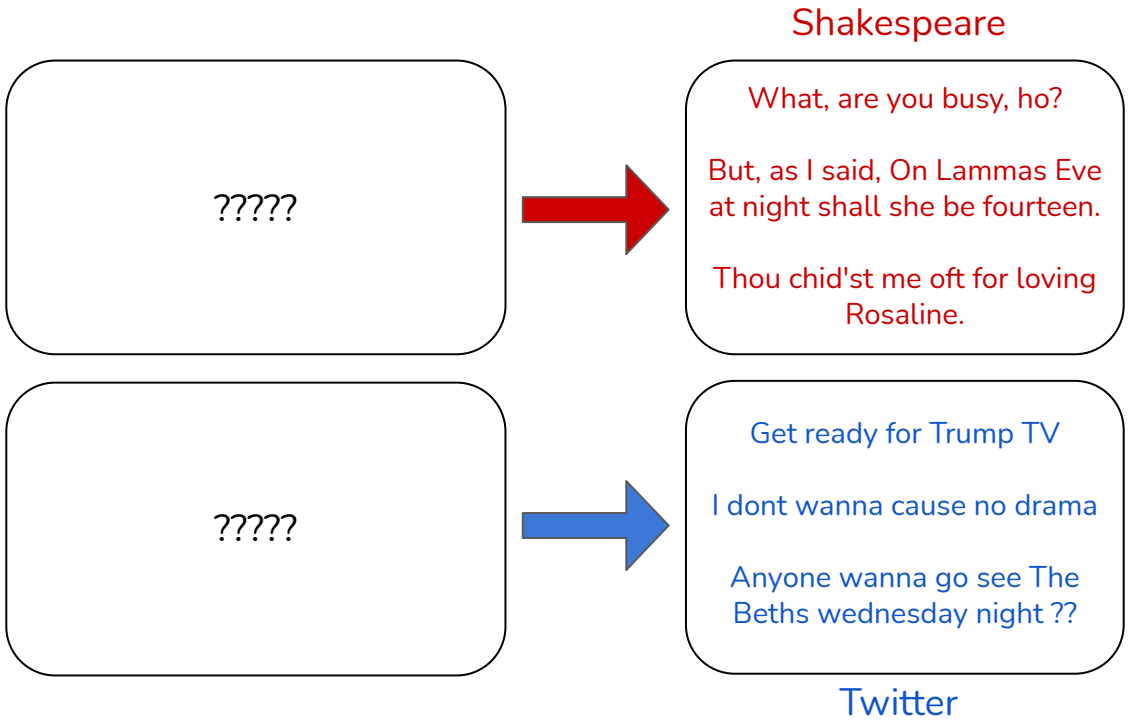
New corpus of **15M** sentences, **11** diverse styles (Tweets, Bible, Poetry, speech transcripts etc.)



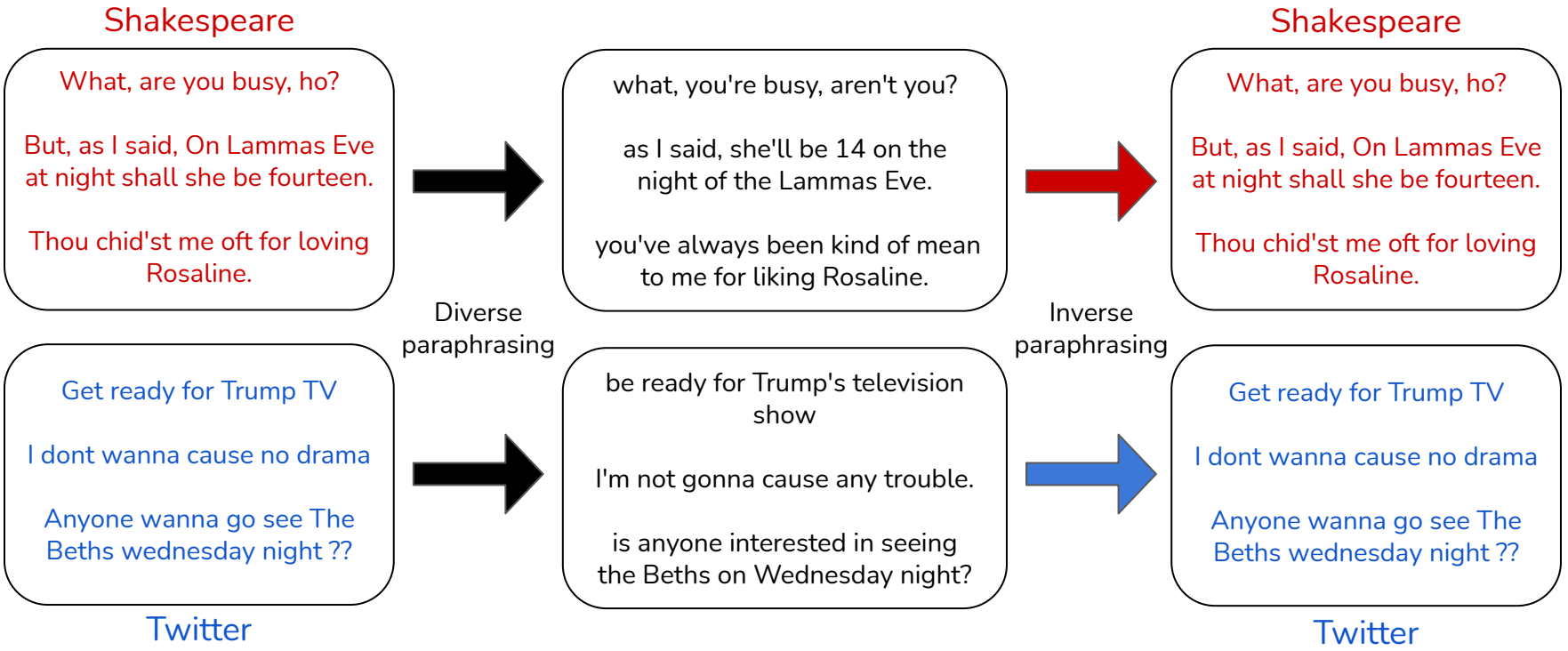
Transfer style while preserving semantics?



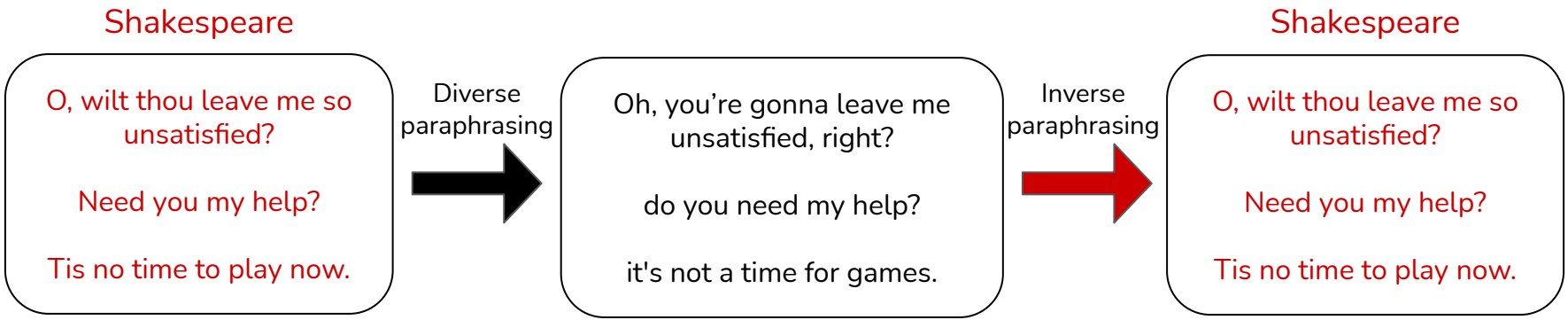
No access to paired data



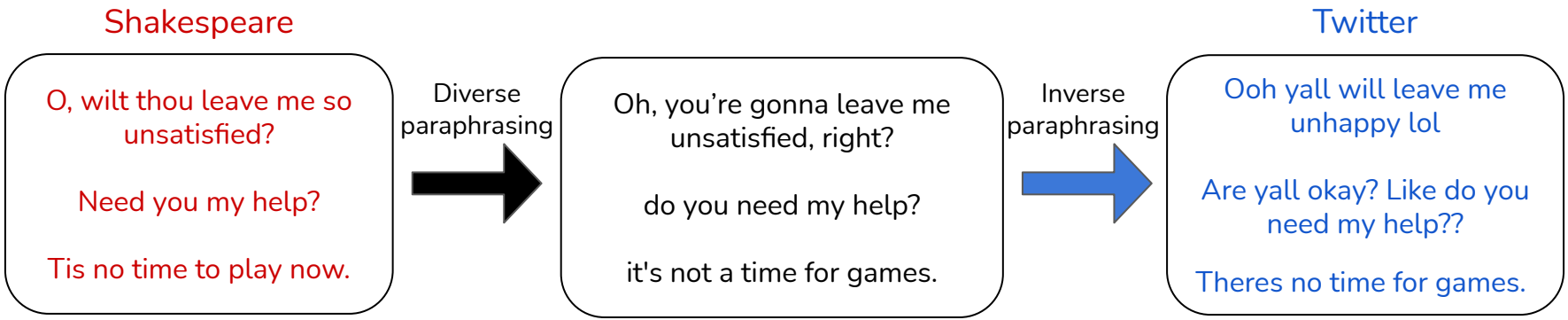
Create pseudo parallel data using **diverse** paraphraser



Swap inverse paraphraser during inference for style transfer!



Swap inverse paraphraser during inference for style transfer!

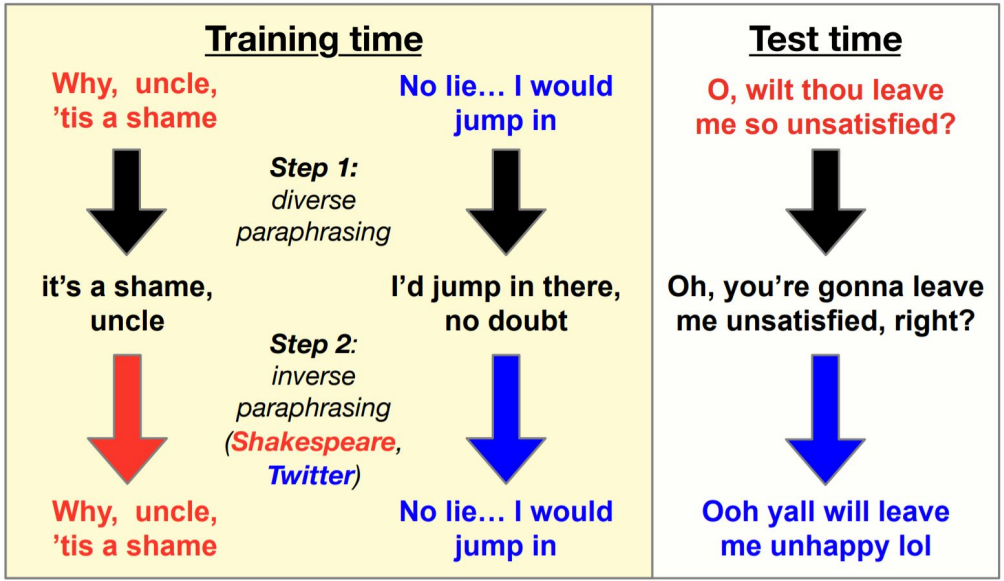


How do they build the diverse paraphraser?

- Trained on back-translated data from ParaNMT-50M

- Aggressively filter ParaNMT-50M (down to 75K pairs) for similarity and diversity
- GPT-2-Large

STRAP: Style Transfer via Paraphrasing

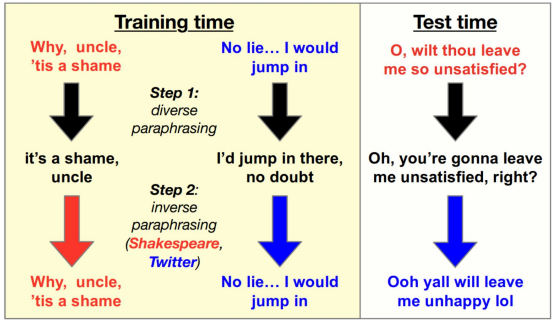


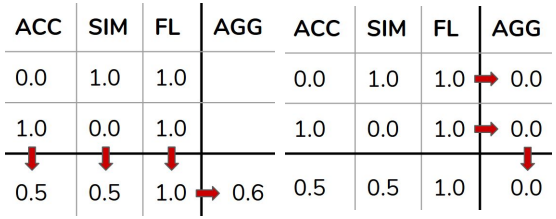
All models are trained by fine-tuning GPT2-large
This improves fluency and generalization

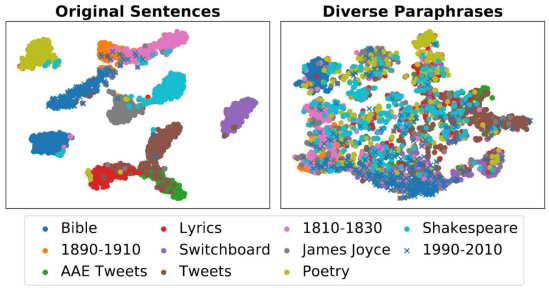
Simple new **state-of-the-art unsupervised** algorithm, models style transfer as controlled paraphrase generation

23-paper survey of evaluation methods, **improvements**

New corpus of **15M** sentences, **11** diverse styles (Tweets, Bible, Poetry, speech transcripts etc.)







Evaluation - independent metrics

(input, target style) → output

Metric	Description	Method
Accuracy	Does the output respect the target style?	Accuracy of style classifier (RB-L)

Evaluation - independent metrics

(input, target style) → output

Metric	Description	Method
Accuracy	Does the output respect the target style?	Accuracy of style classifier (RB-L)
Similarity	Do the input and output share semantics?	Semantic textual similarity benchmark

Evaluation - independent metrics

(input, target style) → output

Metric	Description	Method
Accuracy	Does the output respect the target style?	Accuracy of style classifier (RB-L)
Similarity	Do the input and output share semantics?	Semantic textual similarity benchmark
Fluency	Is the output grammatically correct?	Grammatical acceptability judgments (CoLA)

Evaluation - independent metrics and aggregation

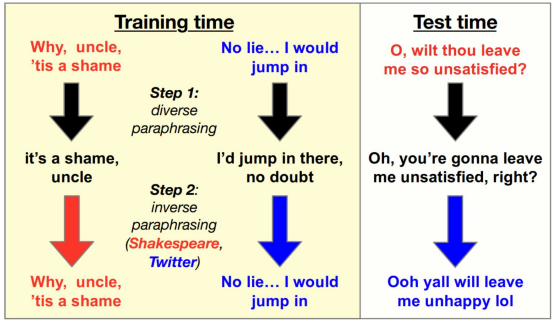
(input, target style) → output

Metric	Description	Method
Accuracy	Does the output respect the target style?	Accuracy of style classifier(RB-L)
Similarity	Do the input and output share semantics?	Semantic textual similarity benchmark
Fluency	Is the output grammatically correct?	Grammatical acceptability judgments (CoLA)

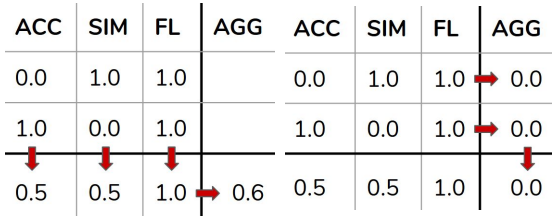
Aggregation: Combine these 3 metrics to a single score

$$J(\mathrm{ACC}, \mathrm{SIM}, \mathrm{FL}) = \sum_{x \in \mathbf{X}} \frac{\mathrm{ACC}(x) \cdot \mathrm{SIM}(x) \cdot \mathrm{FL}(x)}{|\mathbf{X}|}$$

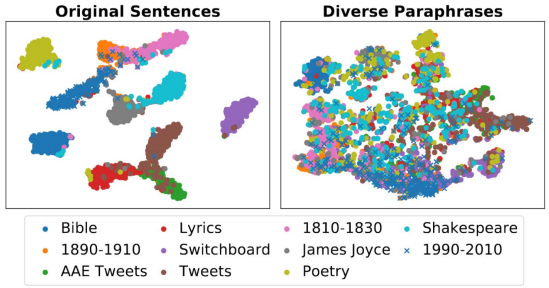
Simple new **state-of-the-art unsupervised** algorithm, models style transfer as controlled paraphrase generation



23-paper survey of evaluation methods, **improvements**



New corpus of **15M** sentences, **11** diverse styles (Tweets, Bible, Poetry, speech transcripts etc.)



Current benchmarks are toyish settings

- carefully curated to remove noisy text
- only two styles
- unpaired, but same content distribution

Corpus of Diverse Styles (CDS)

Style	Size	Style	Size
Shakespeare	27.5K	Lyrics	5.1M
James Joyce	41.2K	1810-1830	216.0K
English Tweets	5.2M	1890-1910	1.3M
AAE Tweets	732.3K	1990-2010	2.0M
Romantic Poetry	29.8K	Bible	34.8K
Switchboard	148.8K		

- 15 million sentences with minimal preprocessing
- 11 diverse styles
- Authors, Tweets, Poetry, Speech Transcripts, Historical Fiction and Biblical Text

Examples from CDS

What, are you busy, ho?

But, as I said, On Lammas Eve at night shall she be fourteen.

Thou chid'st me oft for loving Rosaline.

Speakest thou from thy heart?

.....

Shakespeare

if y- you know instead of

well hi i guess uh

and uh cranberry sauce i- i could eat just that and be satisfied

yeah i think i mean that's pretty i- that's a pretty important uh

.....

Speech Transcripts

Get ready for Trump TV

I dont wanna cause no drama

Anyone wanna go see The Beths wednesday night ??

Enjoy the week ahead! #YoungDrivers #Monday

.....

Tweets

On this poor being all depends,

Spurning nature, torturing art;

But weep, and weep, that they were born so fair?

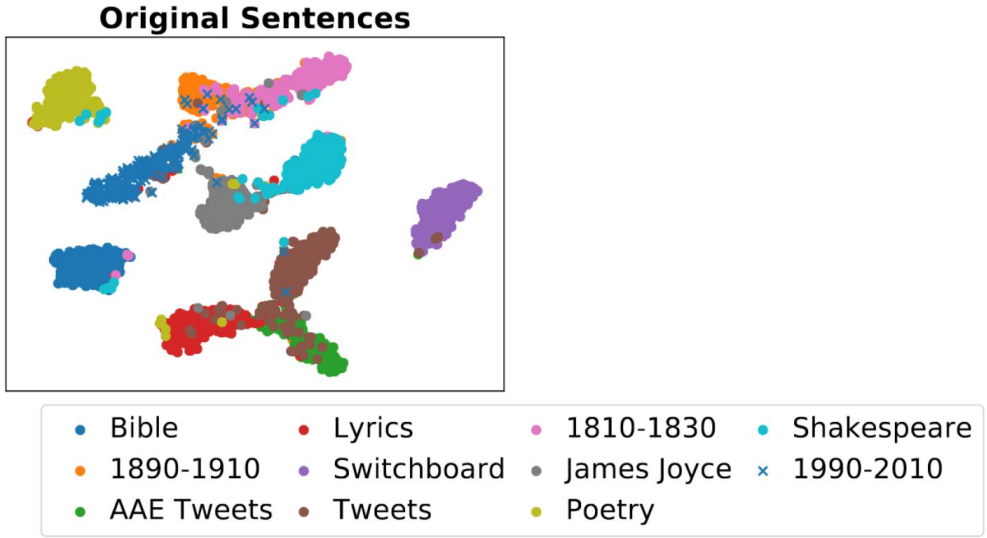
There woos no home, nor hope, nor life, save what is here.

.....

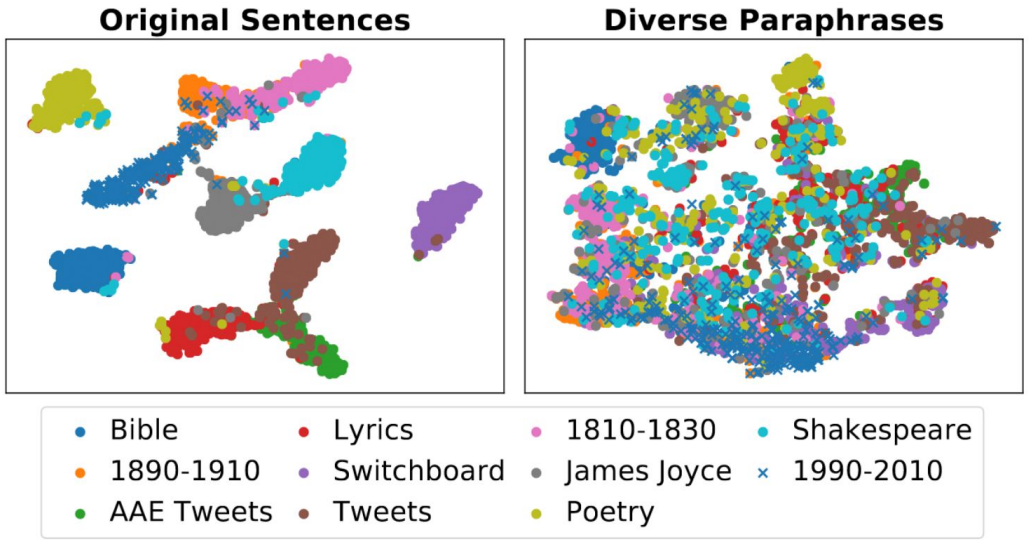
Romantic Poetry

completely different content distributions, diverse vocabularies

What does diverse paraphrasing do to CDS?

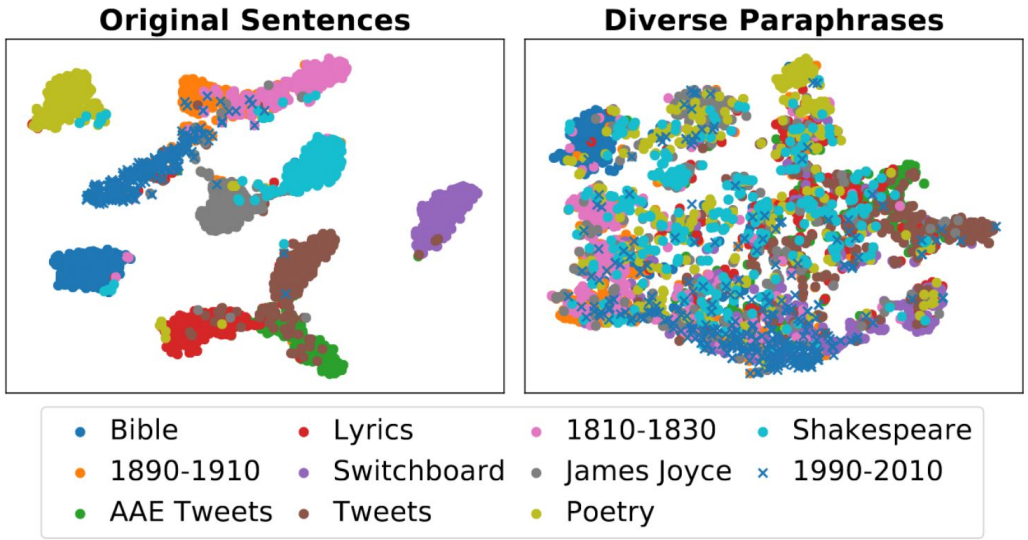


What does diverse paraphrasing do to CDS?



Distortion of clusters formed by dense feature vectors

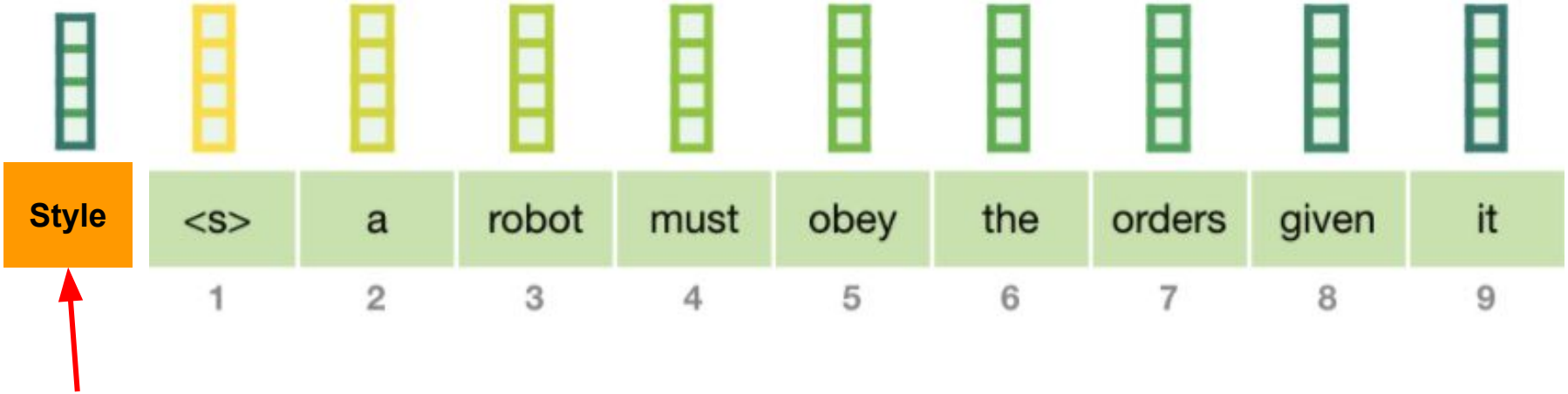
What does diverse paraphrasing do to CDS?



11-way accuracy drops from 88.9% → 42.5%

(51.5% with non-diverse paraphrasing)

Style Controlling



The style indicator vector is a vector that average all [CLS] vectors of training samples in the target style.

Direction	Input	Output
Tweet → Shak.	yall kissing before marriage?	And you kiss'd before your nuptial?
Lyrics → AAE	It's a good thing you don't have bus fare	It's a good thing u aint gettin no ticket
Tweet → Poetry	Its so disrespectful I hate ittttttt	For 'tis so rude, I have a sick regard,
Shak. → Tweet	Need you my help?	Are yall okay? Like do you need my help??
AAE → Shak.	Aint got nooooo holiday spirit frfr	No spirit of this holiday, gentle man.
Poetry → Bible	For the foul thief is just at your gate.	the wicked thief art at thy door.
1990. → Tweet	Now I can't talk to him about it.	I dont wanna talk to him abt that
1990. → Poetry	I don't know why, but I sensed there was something wrong.	Something felt wrong; I knew not why,
Tweet → Bible	when ur going thru an episode and cant cope anymore	when thou art in an episode, and cannot stand it any more;
1890. → AAE	I was just thinking the same thing that you have expressed.	u said the same thing i thought lol
1990. → 1810.	He was being terrorized into making a statement by the same means as the other so-called "witnesses."	Terror had been employed in the same manner with the other witnesses, to compel him to make a declaration.
AAE → Shak.	If I got a dollar every time one of my friends told me they hate me, I'd be rich	I would have been rich, had I but a dollar for every friend that hath said they hate me.
Joyce → Bible	I appeal for clemency in the name of the most sacred word our vocal organs have ever been called upon	I beseech thee in the name of the most holy word which is in our lips, forgive us our trespasses.

Results

STRAP outperforms prior work on automatic evaluation

Formal English ⇌ Informal English

Model	ACC	SIM	FL	S. AGG
UNMT (Lample et al. ICLR 2019)	**78.5**	49.1	52.5	20.0
DLSM (He et al. ICLR 2020)	78.0	47.7	53.7	18.6
STRAP (Ours)	67.7	**72.5**	**90.4**	**45.5**

STRAP outperforms prior work on automatic evaluation

Shakespearean English ⇌ Modern English

Model	ACC	SIM	FL	S. AGG
UNMT (Lample et al. ICLR 2019)	70.5	37.5	49.6	14.6
DLSM (He et al. ICLR 2020)	71.1	43.5	49.4	16.3
STRAP (Ours)	71.7	**56.4**	**85.2**	**34.7**

Consistent trends even on human evaluation

	Formality Transfer			Shakespeare		
Model	ACC	SIM	S. AGG	ACC	SIM	S. AGG
UNMT (Lample et al. ICLR 2019)	77.3	22.7	7.3	69.3	20.7	7.3
DLSM (He et al. ICLR 2020)	78.0	24.0	10.0	65.3	37.3	9.3
STRAP (Ours)	71.3	**76.0**	**41.3**	70.7	**79.3**	**47.3**

Ablation studies

Dataset	Model	ACC	SIM	FL	J(A,S,F)
Form.	STRAP	67.7	72.5	90.4	45.5
	– Inf. PP	27.5	78.5	88.2	20.7
	– Mult. PP	63.1	72.0	90.8	42.3
	– Div. PP	61.2	79.5	88.7	43.8
	– GPT2	84.6	43.8	61.7	23.1
	GPT2-md	71.0	70.7	88.6	45.8
	GPT2-sm	69.1	68.6	87.6	42.9
Shak.	STRAP	71.7	56.4	85.2	34.7
	– Inf. PP	40.1	66.1	76.3	23.3
	– Mult. PP	45.9	56.5	91.1	24.8
	– Div. PP	49.7	64.4	82.9	28.2
	– GPT2	75.6	26.7	66.9	13.6
	GPT2-md	73.4	54.0	86.4	34.3
	GPT2-sm	68.0	53.2	84.6	31.5

Table 3: Ablation study using automatic metrics on the Formality (Form.) and Shakespeare (Shak.) datasets.

Style Transfer on CDS

Shakespeare ⇌ Tweets

Model	ACC	SIM	FL	S. AGG
UNMT (Lample et al. ICLR 2019)	**76.7**	20.6	37.7	4.4
DLSM (He et al. ICLR 2020)	64.2	19.6	33.1	2.0
STRAP (Ours)	43.2	54.5	68.3	**13.9**

Prior models struggle on this task, often generating arbitrary attribute-specific text