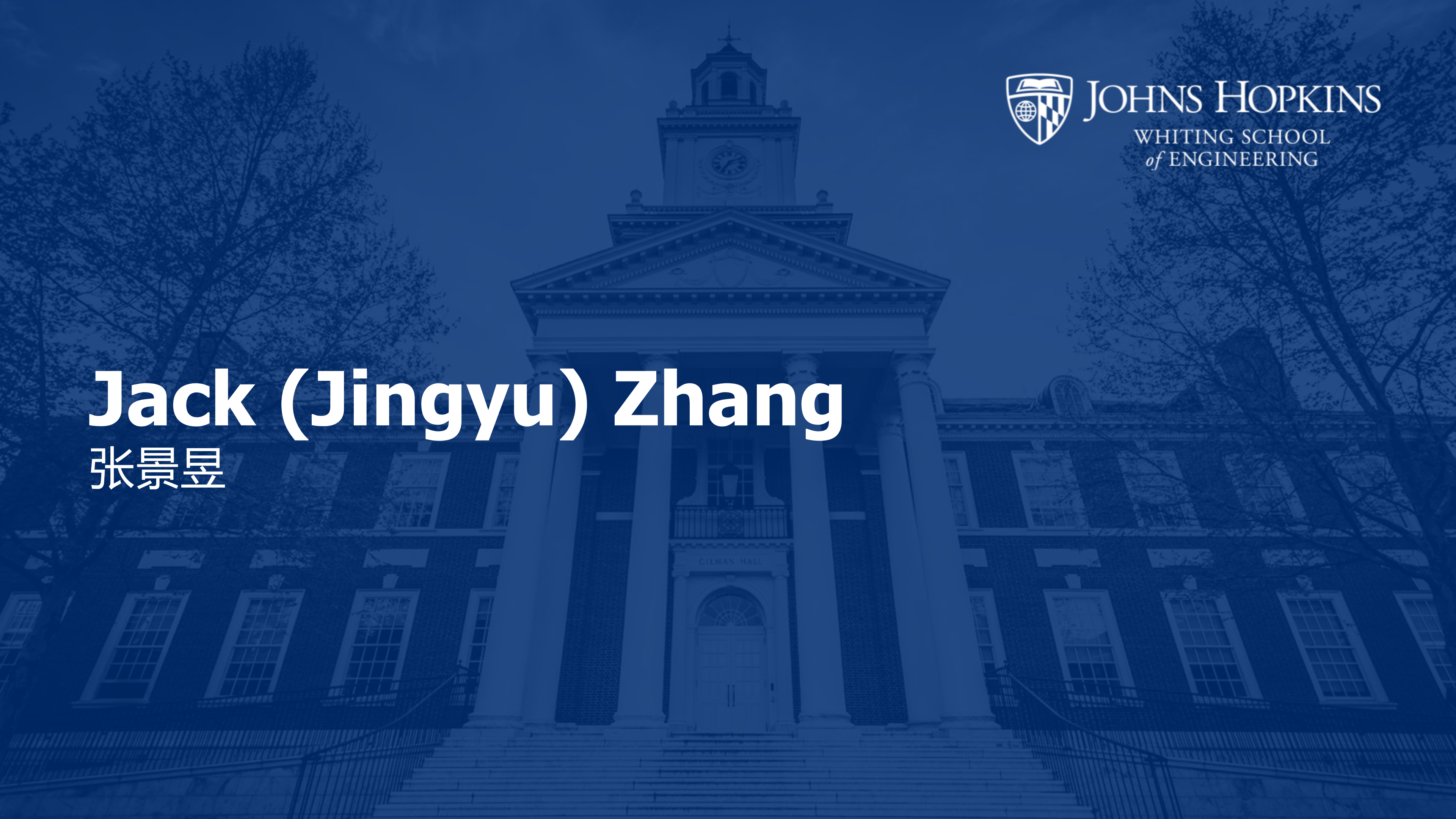


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Jack (Jingyu) Zhang

张景昱

About Me

- 4th year undergrad at JHU (triple major in CS, Math, Applied Math)
- Have been conducting research in NLP since sophomore year
- Before NLP: C++ Dev at ByteDance
- In the past, I have worked on
  - Natural Language Generation
  - Computational Social Science
- Interested in exploring new directions!

Publications & Preprints

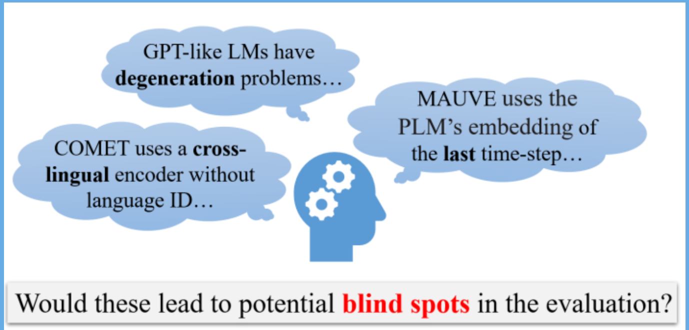
NLG related:

- Tianxing He*, **Jingyu Zhang***, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, Yulia Tsvetkov. On the Blind Spots of Model-Based Evaluation Metrics for Text Generation. *Submitted to ACL 2023*.
- **Jingyu Zhang**, James Glass, Tianxing He. PCFG-based Natural Language Interface Improves Generalization for Controlled Text Generation. *Submitted to ACL 2023*. Preliminary version accepted at *2nd Workshop on Efficient Natural Language and Speech Processing (ENLSP), NeurIPS 2022*. ***Best Paper Award***.

CSS related:

- **Jingyu Zhang**, Alexandra DeLucia, Chenyu Zhang, Mark Dredze. Geo-Seq2seq: Twitter User Geolocation on Noisy Data through Sequence to Sequence Learning. *Submitted to ACL 2023*.
- **Jingyu Zhang**, Alexandra DeLucia, Mark Dredze. Changes in Tweet Geolocation over Time: A Study with Carmen 2.0. *Proceedings of the 8th Workshop on Noisy User-generated Text (W-NUT), COLING 2022*.
- Abhinav Chinta*, **Jingyu Zhang***, Alexandra DeLucia, Anna L. Buzcak, Mark Dredze. Study of Manifestation of Civil Unrest on Twitter. *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT), EMNLP 2021*.

*Equal Contribution


GPT-like LMs have **degeneration** problems…
MAUVE uses the PLM's embedding of the **last** time-step…
COMET uses a **cross-lingual** encoder without language ID…

Would these lead to potential **blind spots** in the evaluation?

On the Blind Spots of Model-Based Evaluation Metrics for Text Generation

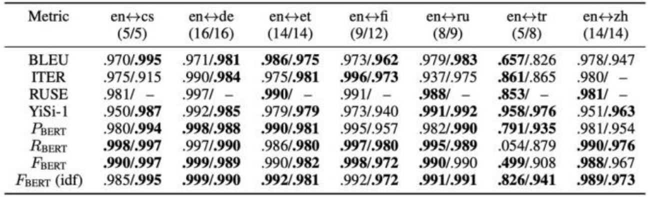
Tianxing He*, **Jingyu Zhang***, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, Yulia Tsvetkov.

Submitted to ACL 2023

Motivation

- Recently, a series of PLM-based metrics (BERTScore, MAUVE, etc.) is shown to correlate well with human annotations.
- The correlation scores are not very informative in the sense that they do not point out which aspect of the metric needs to be improved.

example from the BERTScore paper ->

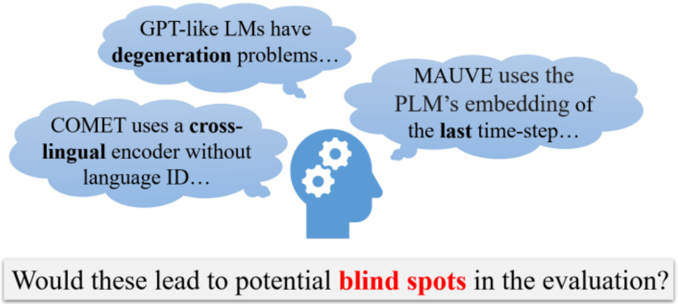


Metric	en↔cs (5/5)	en↔de (16/16)	en↔et (14/14)	en↔fi (9/12)	en↔ru (8/9)	en↔tr (5/8)	en↔zh (14/14)
BLEU	.970/**.995**	.971/**.981**	**.986/.975**	.973/**.962**	.979/**.983**	**.657**/.826	.978/.947
ITER	.975/.915	.990/**.984**	.975/**.981**	**.996/.973**	.937/.975	**.861**/.865	.980/ –
RUSE	.981/ –	.997/ –	**.990/** –	.991/ –	**.988/** –	**.853/** –	**.981/** –
YiSi-1	.950/**.987**	.992/**.985**	.979/**.979**	.973/.940	**.991/.992**	**.958/.976**	.951/**.963**
P_{BERT}	.980/**.994**	**.998/.988**	**.990/.981**	.995/.957	.982/**.990**	**.791/.935**	.981/.954
R_{BERT}	**.998/.997**	.997/**.990**	.986/**.980**	**.997/.980**	**.995/.989**	.054/.879	**.990/.976**
F_{BERT}	**.990/.997**	**.999/.989**	.990/**.982**	**.998/.972**	**.990**/.990	**.499**/.908	**.988**/.967
F_{BERT} (idf)	.985/**.995**	**.999/.990**	**.992/.981**	.992/**.972**	**.991/.991**	**.826/.941**	**.989/.973**

Table 1: Absolute Pearson correlations with system-level human judgments on WMT18. For each language pair, the left number is the to-English correlation, and the right is the from-English. We bold correlations of metrics not significantly outperformed by any other metric under Williams Test for that language pair and direction. The numbers in parenthesis are the number of systems used for each language pair and direction.

Motivation

- On the other hand, the PLMs are not perfect. Known issues: degeneration, various inducive bias, insensitivity to certain linguistic phenomenon (e.g., negation), etc. These imperfection could lead to undesirable behavior of the induced metrics.
- Certain design choices can also lead to potential problems in evaluation



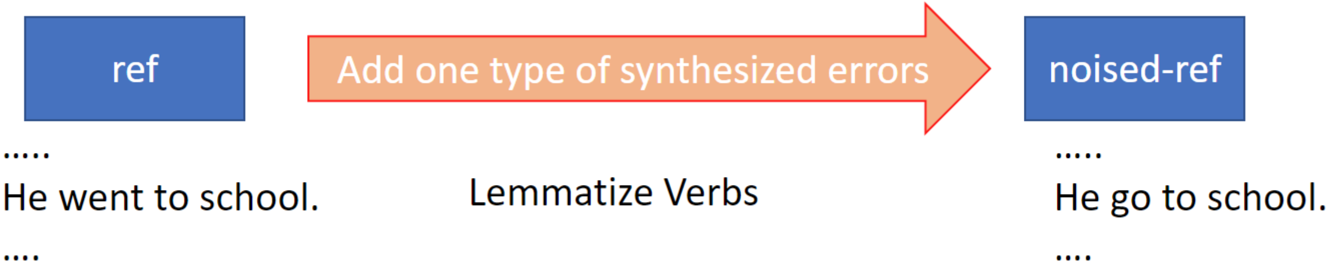
GPT-like LMs have **degeneration** problems…

MAUVE uses the PLM's embedding of the **last** time-step…

COMET uses a **cross-lingual** encoder without language ID…

Would these lead to potential **blind spots** in the evaluation?

Protocol for “Stress Test”



ref

Add one type of synthesized errors

noised-ref

.....

He went to school.

....

Lemmatize Verbs

.....

He go to school.

....

- We conduct a systematic series of stress tests for PLM-based NLG metrics
- In each test, we design and synthesize one type of textual error and check whether it results in a commensurate drop in the metric score

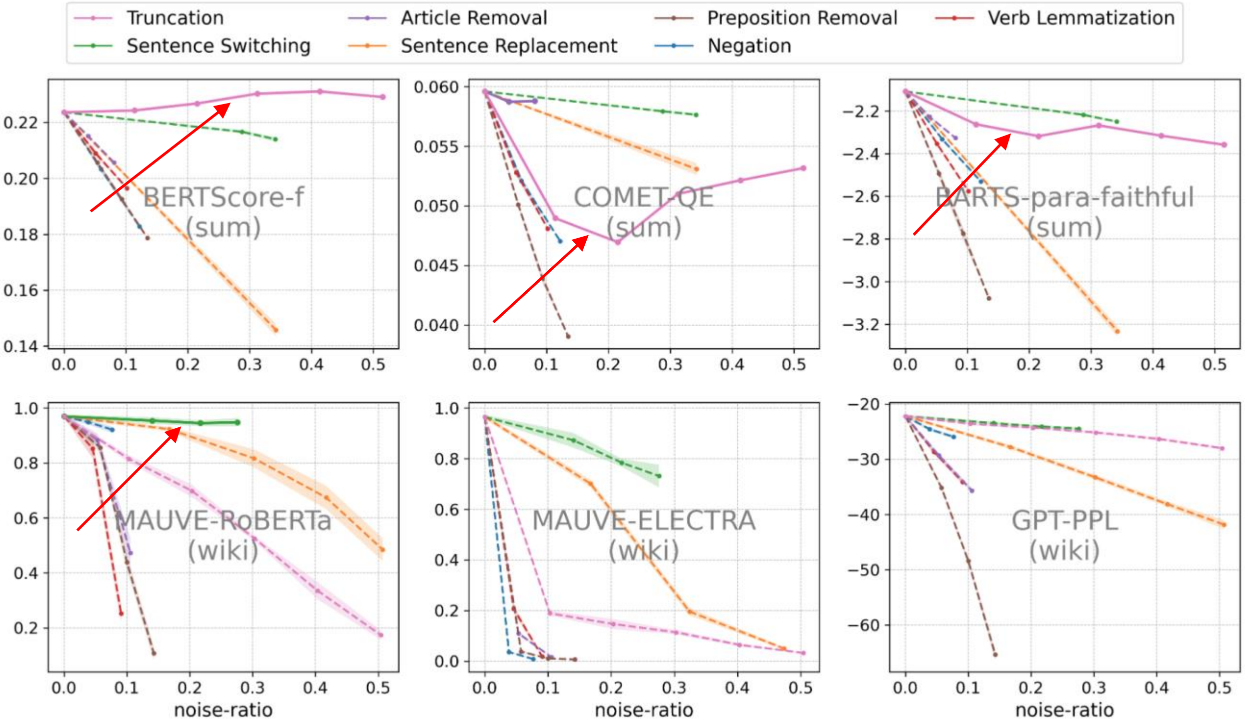
Rank-Based Tests

- As we add more noise, metric score should decrease more…
- Quantify the amount of noise using “noise-ratio”, based on Levenshtein distance:

$$\frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}\frac{\text{Levenshtein}(h', h)}{\text{len}(h)}$$

where H is the set of gold hypotheses, h’ is the noised hypothesis.

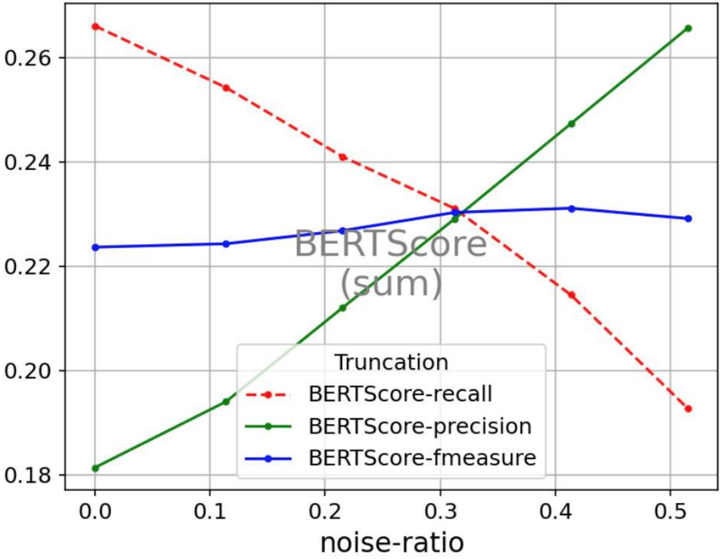
Rank-Based Tests



A subset of results

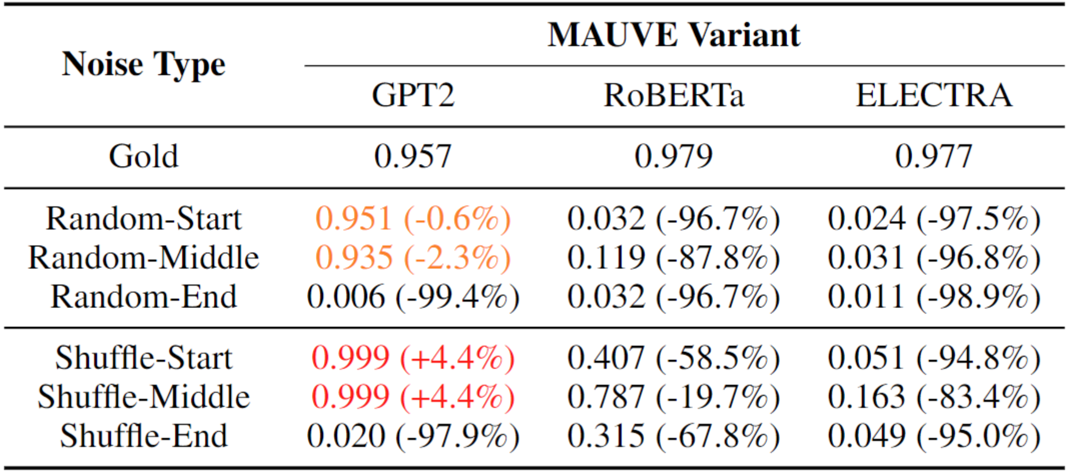
Example "blind spots"

- BERTScore (Zhang et al., 2020) is confused by synthetic truncation error on summarization data



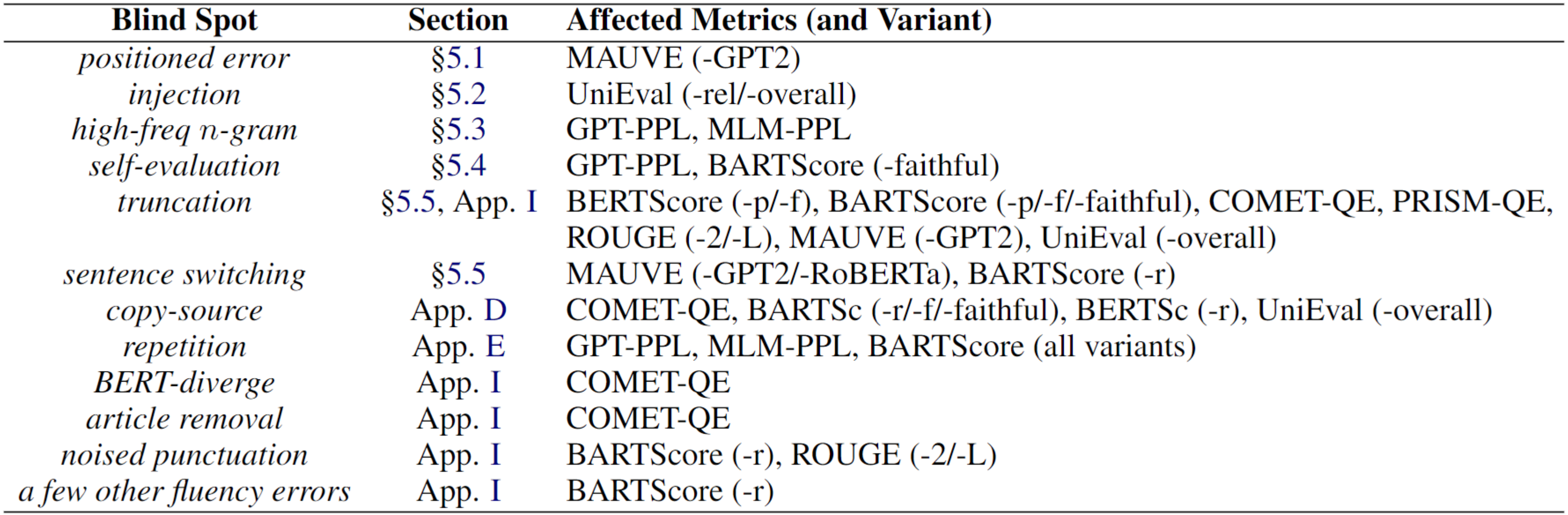
Example “blind spots”

- MAUVE (Pillutla et al., 2021) based on GPT-2 features ignores synthetic errors in the middle or start of the hypothesis.



Noise Type	MAUVE Variant		
	GPT2	RoBERTa	ELECTRA
Gold	0.957	0.979	0.977
Random-Start	0.951 (-0.6%)	0.032 (-96.7%)	0.024 (-97.5%)
Random-Middle	0.935 (-2.3%)	0.119 (-87.8%)	0.031 (-96.8%)
Random-End	0.006 (-99.4%)	0.032 (-96.7%)	0.011 (-98.9%)
Shuffle-Start	0.999 (+4.4%)	0.407 (-58.5%)	0.051 (-94.8%)
Shuffle-Middle	0.999 (+4.4%)	0.787 (-19.7%)	0.163 (-83.4%)
Shuffle-End	0.020 (-97.9%)	0.315 (-67.8%)	0.049 (-95.0%)

A Grand Summary of Blind Spots



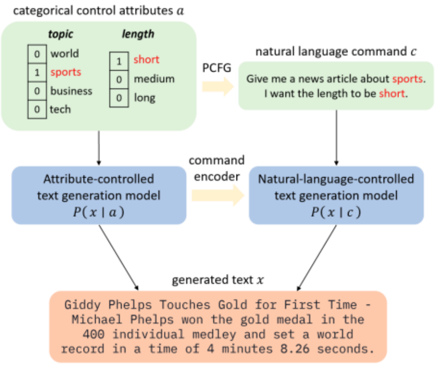
Blind Spot	Section	Affected Metrics (and Variant)
positioned error	§5.1	MAUVE (-GPT2)
injection	§5.2	UniEval (-rel/-overall)
high-freq n-gram	§5.3	GPT-PPL, MLM-PPL
self-evaluation	§5.4	GPT-PPL, BARTScore (-faithful)
truncation	§5.5, App. I	BERTScore (-p/-f), BARTScore (-p/-f/-faithful), COMET-QE, PRISM-QE, ROUGE (-2/-L), MAUVE (-GPT2), UniEval (-overall)
sentence switching	§5.5	MAUVE (-GPT2/-RoBERTa), BARTScore (-r)
copy-source	App. D	COMET-QE, BARTSc (-r/-f/-faithful), BERTSc (-r), UniEval (-overall)
repetition	App. E	GPT-PPL, MLM-PPL, BARTScore (all variants)
BERT-diverge	App. I	COMET-QE
article removal	App. I	COMET-QE
noised punctuation	App. I	BARTScore (-r), ROUGE (-2/-L)
a few other fluency errors	App. I	BARTScore (-r)

Conclusion

- Using pretrained language models for NLG metrics is a **double-edged sword**!
  - Benefit: powerful representations
  - Danger: black-box nature of PLMs may cause unexpected behavior
- We show that stress tests, **complimentary** to the standard human correlation tests, are powerful tools to cover corner cases and point out aspects where the metric could improve.
- A call for caution for both metric users and metric developers
- More generally, a better understanding of PLM itself is needed

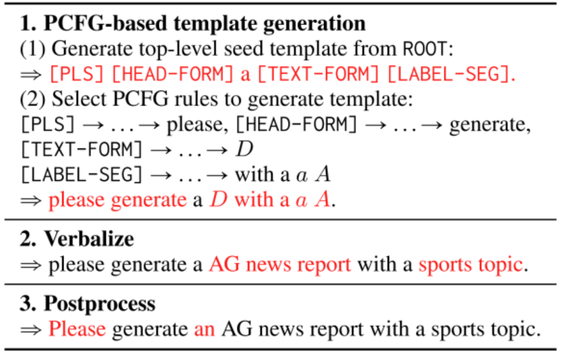
Other Work

- **PCFG-based Natural Language Interface Improves Generalization for Controlled Text Generation**: enable controllable text generation methods to generalize to unseen control attributes (e.g., topics, sentiment, formality) by crafting a PCFG to embed the control attributes into natural language commands


categorical control attributes a
topic
0 world
1 sports
0 business
0 tech
length
1 short
0 medium
0 long
PCFG
natural language command c
Give me a news article about sports.
I want the length to be short.
command encoder
Attribute-controlled text generation model $P(x \mid a)$
Natural-language-controlled text generation model $P(x \mid c)$
generated text x

Giddy Phelps Touches Gold for First Time - Michael Phelps won the gold medal in the 400 individual medley and set a world record in a time of 4 minutes 8.26 seconds.

Figure 1: We use PCFG to embed categorical control attributes into natural language command. Correspondingly, we propose generation models that take command as input.



1. PCFG-based template generation

(1) Generate top-level seed template from ROOT:

⇒ [PLS] [HEAD-FORM] a [TEXT-FORM] [LABEL-SEG].

(2) Select PCFG rules to generate template:

[PLS] → . . . → please, [HEAD-FORM] → . . . → generate,
[TEXT-FORM] → . . . → D
[LABEL-SEG] → . . . → with a a A

⇒ please generate a D with a a A.

2. Verbalize

⇒ please generate a AG news report with a sports topic.

3. Postprocess

⇒ Please generate an AG news report with a sports topic.

Table 1: Examples of PCFG command generation. ROOT is the PCFG start symbol. Newly replaced segments are highlighted in red. In step 1.(2), we omit intermediate PCFG expansions to "→ . . . →".

Other Work

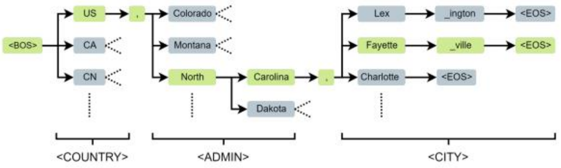


Figure 3: Excerpt from the “reversed" decoding trie built from the Carmen location database. The output sequence is constrained at each overarching step to `<BOS>` → `<COUNTRY>` → ... `<EOS>`. At each sub-step, the generated tokens are constrained to valid subwords, or those present in the location database at that step.

- **Twitter User Geolocation on Noisy Data through Sequence to Sequence Learning**: Rewrites noisy, multilingual user-provided location strings into structured location names using
  - mT5-based multilingual geolocation name transducer
  - Trie-based constrained decoding

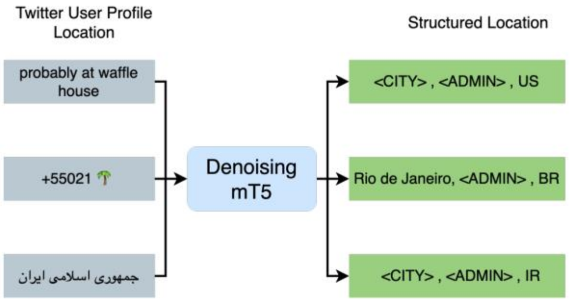


Figure 1: Geolocation of a user profile location to structured location. For example, GEO-SEQ2SEQ correctly maps “waffle house” (a US-based restaurant) to the US, a zip code to Brazil, and the Farsi name for Iran to Iran.

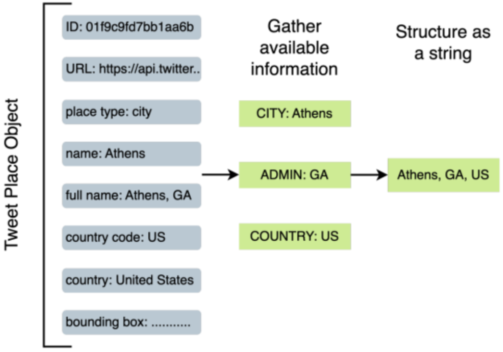
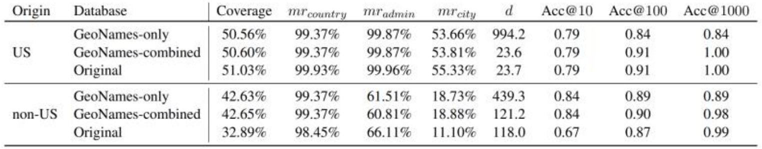


Figure 2: Ground truth label created from the tweet place objects. Each ground truth string is of the form “`<CITY>,<ADMIN>,<COUNTRY>`." The special tokens are left as-is when information is not available or does not apply.

Other Work

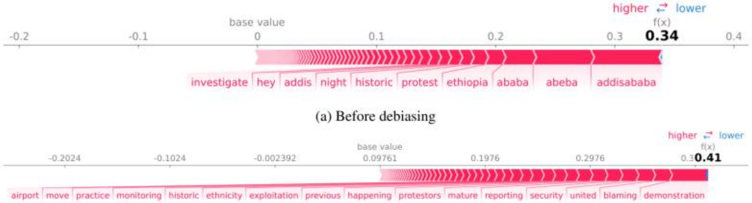
- **Changes in Tweet Geolocation over Time: A Study with Carmen 2.0**: analyze the performance of a Twitter geotagger under different languages, countries, and time



Origin	Database	Coverage	$mr_{country}$	mr_{admin}	mr_{city}	d	Acc@10	Acc@100	Acc@1000
US	GeoNames-only	50.56%	99.37%	99.87%	53.66%	994.2	0.79	0.84	0.84
	GeoNames-combined	50.60%	99.37%	99.87%	53.81%	23.6	0.79	0.91	1.00
	Original	51.03%	99.93%	99.96%	55.33%	23.7	0.79	0.91	1.00
non-US	GeoNames-only	42.63%	99.37%	61.51%	18.73%	439.3	0.84	0.89	0.89
	GeoNames-combined	42.65%	99.37%	60.81%	18.88%	121.2	0.84	0.90	0.98
	Original	32.89%	98.45%	66.11%	11.10%	118.0	0.67	0.87	0.99

Table 3: Ablation over Carmen location database and performance on tweets originating from and outside of the United States (US). Evaluated on TWITTER-GLOBAL. "Acc@K" represents the ratio of tweets predicted within K miles of the ground truth. Higher values are best for all metrics except distance (d).

- **Study of Manifestation of Civil Unrest on Twitter**: using explainability tools (SHAP) to identify important feature learned by ML models to discover words indicative of civil unrest that generalized across countries



(a) Before debiasing

(b) After debiasing

Figure 7: Individual SHAP output on days of the Burayu Massacre protests in Ethiopia before and after country debiasing (i.e., removing location-specific words). While the overall model performance dropped 0.05 F1, the most important features became more generalizable and informative.

Thank you!

- Personal website: jackz.io
- Questions?