Automatic Evaluation for Grammatical Error Correction in the Era of Large Language Models

Hitotsubashi University
Graduate School of Social Data Science
Mamoru Komachi
<mamoru.komachi@r.hit-u.ac.jp>

Writing in English: A challenge for non-native speakers

- Natural language processing techniques have been widely accepted

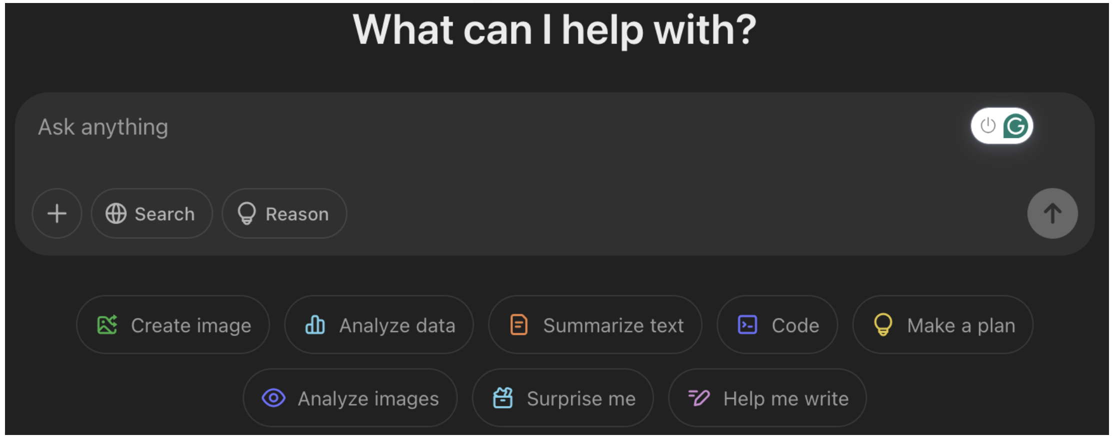
Microsoft Research | ESL Assistant

grammarly

GINGER

Advancements in deep learning and their impact on English writing assistance

- Large language models (LLMs) can help!
- LLMs can be used as an automatic evaluator



Two research questions for GEC evaluation

Are the existing datasets adequate in the era of deep learning?

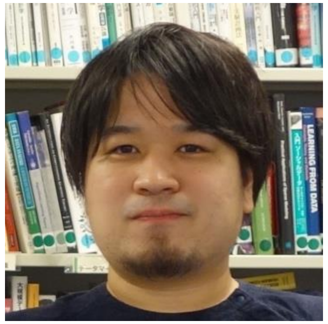
→Revisiting meta-evaluation (evaluation of evaluations) for grammatical error correction

Can LLMs be used to evaluate grammatical error correction?

→Application of LLMs for evaluation of grammatical error correction (LLM-as-a-judge)

Revisiting Meta-evaluation for Grammatical Error Correction (Transactions of the Association for Computational Linguistics 2024)

Joint work with Masamune Kobayashi and Masato Mita



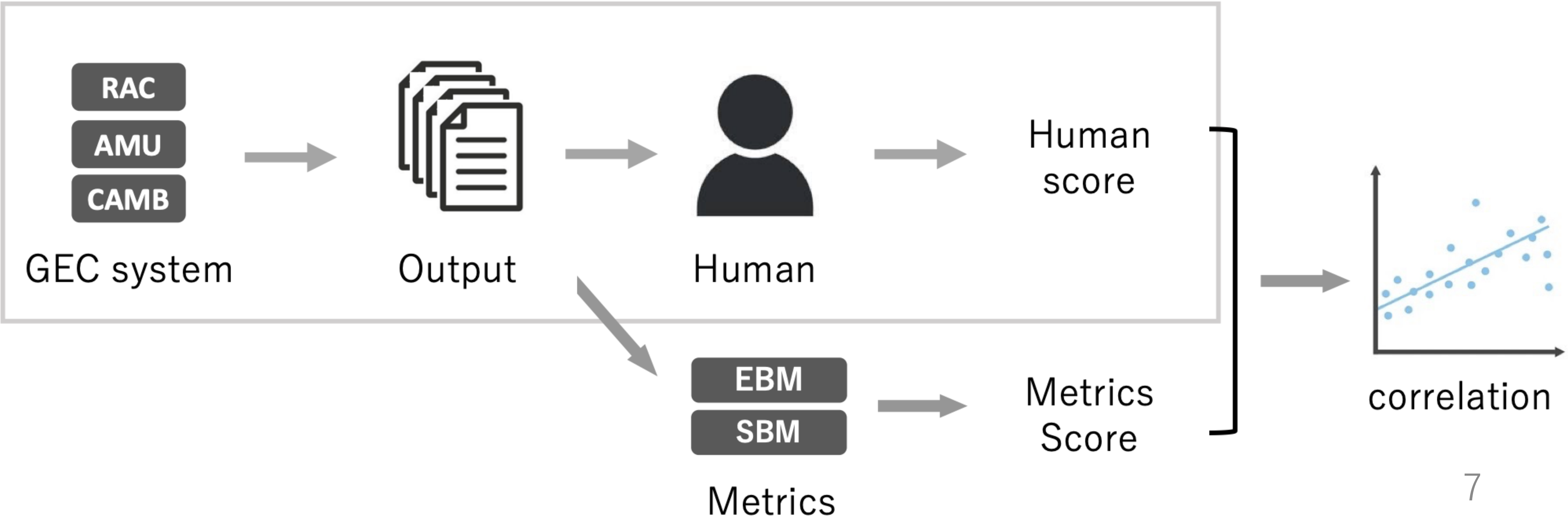
Automatic evaluation of GEC: edit-based and sentence-based metrics

- Two types of grammatical error correction (GEC) evaluation metrics based on (human) evaluation granularity

Edit-Based Metrics (EBMs)	Sentence-Based Metrics (SBMs)
• Evaluate only each edit	• Evaluate the quality as a sentence
I [go → went] to Tokyo [yestaday → yesterday].	I went to Tokyo yesterday.
Score A, Score B	Score Y
⬇	⬇
Evaluation score X (=A+B)	Evaluation score Y

Meta-evaluation (evaluation of evaluations) of GEC using human judgment

- Grundkiewicz+ (2015) dataset (GJG15) is the most well-known dataset for meta-evaluation of GEC



Major issues in previous evaluation methods and their meta-evaluation

1. Discrepancy in evaluation granularity: Human evaluators consider a broader context, whereas automatic metrics typically rely on minimal context
2. Human judgment on classical systems: GJG15 conducts human evaluations on traditional systems predating the emergence of deep learning models
3. Impact of outlier systems in meta-evaluation: The presence of outlier systems can influence overall conclusion, particularly when using a single configuration

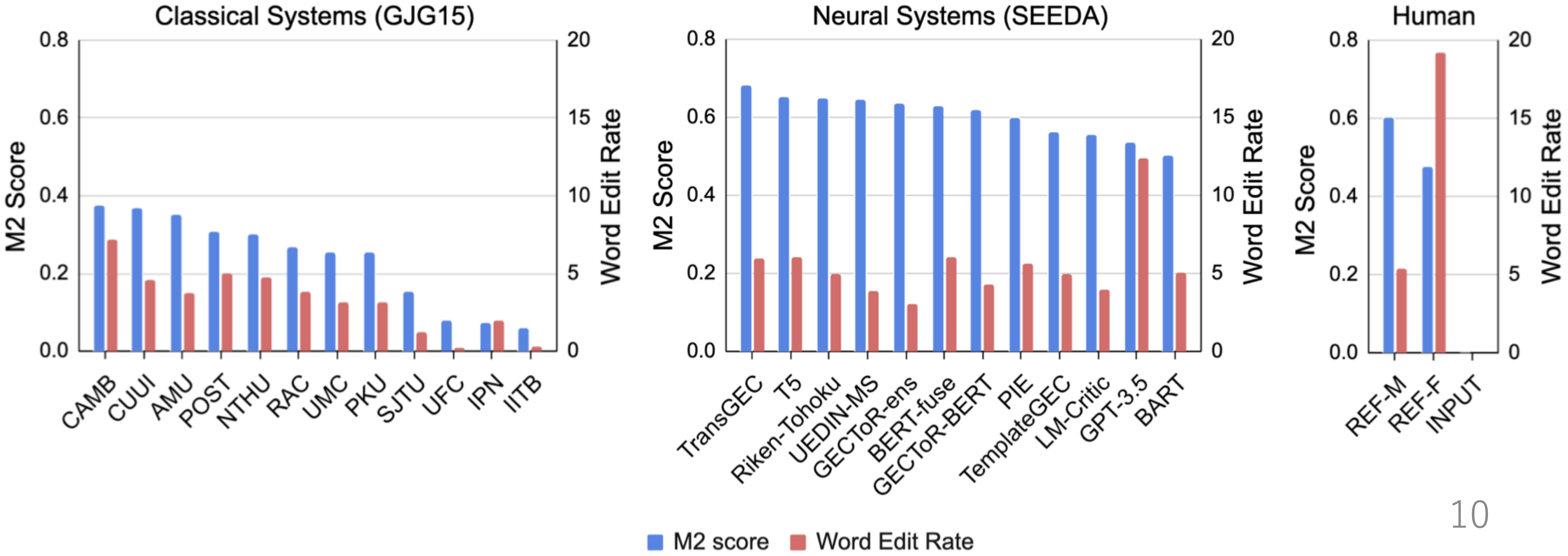
Main contributions of this work

1. Construction of the SEEDA dataset
    - Annotations were conducted at both the edit level and the sentence level
    - Various types of neural systems were annotated

2. Comprehensive meta-evaluation
    - Conducted across a wide range of settings
    - Examines the potential impact of outliers and system variations

High-performance modern GEC systems were chosen as annotation targets

- Neural systems generate more edits and better corrections compared to classical systems included in the GJG15 dataset



Two types of evaluation granularity

Edit-based evaluation

Step 1	**Source:** There is a story of a girl who lives in _ social media world every night in eight years. **Output:** There is a story of a girl who **[lives → alive]** in **[→ the]** social media world every night **[in → for]** eight years.
Step 2	**Source:** There is a story of a girl who **[lives]** in **[_]** social media world every night **[in]** eight years.
Score	$F_{0.5} = 0.67$, Precision = 0.67, Recall = 0.67

Sentence-based evaluation

And both are not what we want since most of us just want to live as normal people .
Surrounded by such concerns , it is very likely that we are distracted to worry about these problems .
It is a concern that will be with us during our whole life , because we will never know when the ''potential bomb' ' will explode .
— Source with context

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst
Surrounded by such concerns , it is very likely that we are too distracted to worry about these problems .
— Correction 1

Best ← Rank 1 Rank 2 Rank 3 Rank 4 Rank 5 → Worst
Surrounded by such concerns , it is very likely that we are distracted to worry about these problems .
— Correction 2

Our dataset has higher inter- and intra-annotator agreement than GJG15

Statistics of our dataset (sentences)

	Unexpanded	Expanded
1	1,777	10,893
2	1,770	11,663
3	1,800	10,988
Total	5,347	33,544

Expanded = unroll system outputs by aggregating pairwise evaluation

Annotator agreement	Value	Degree
Inter- (Edit)	0.28	Fair
Inter- (Sentence)	0.41	Moderate
Intra- (Edit)	0.61	Substantial
Intra- (Sentence)	0.71	Substantial

GPT and T5 can produce corrections equivalent to or better than humans

#	Score	Range	System
1	0.273	1	AMU
2	0.182	2	CAMB
3	0.114	3-4	RAC
	0.105	3-5	CUUI
	0.080	4-5	POST
4	-0.001	6-7	PKU
	-0.022	6-8	UMC
	-0.041	7-10	UFC
	-0.055	8-11	IITB
	-0.062	8-11	INPUT
	-0.074	9-11	SJTU
5	-0.142	12	NTHU
6	-0.358	13	IPN

(a) Sentence-based evaluation in GJG15

#	Score	Range	System
1	0.992	1	REF-F
2	0.743	2	GPT-3.5
3	0.179	3-4	T5
	0.175	3-4	TransGEC
4	0.067	5-6	REF-M
	0.023	5-7	BERT-fuse
	-0.001	6-8	Riken-Tohoku
	-0.034	7-8	PIE
5	-0.163	9-12	LM-Critic
	-0.168	9-12	TemplateGEC
	-0.178	9-12	GECToR-BERT
	-0.179	9-12	UEDIN-MS
6	-0.234	13	GECToR-ens
7	-0.300	14	BART
8	-0.992	15	INPUT

(b) Sentence-based evaluation in SEEDA

#	Score	Range	System
1	0.679	1	REF-F
2	0.583	2	GPT-3.5
3	0.173	3	TransGEC
4	0.097	4-6	T5
	0.078	4-7	REF-M
	0.067	4-7	Riken-Tohoku
	0.064	4-7	BERT-fuse
5	-0.076	8-11	UEDIN-MS
	-0.084	8-11	PIE
	-0.092	8-11	GECToR-BERT
	-0.097	8-11	LM-Critic
6	-0.154	12-12	GECToR-ens
7	-0.211	13-14	TemplateGEC
	-0.231	13-14	BART
8	-0.797	15	INPUT

(c) Edit-based evaluation in SEEDA

Meta-evaluation experiment

Target metrics

- Edit-based: M^2, SentM^2, PT-M^2, ERRANT, SentERRANT, PT-ERRANT, GoToScorer
- Sentence-based: GLEU, Scribendi Score, SOME, IMPARA

Meta-evaluation method

- System-level: correlation with human rankings
- Sentence-level: consistency with pairwise judgment

Aligning the evaluation granularity between human and system improves correlation

	Metric	System-level GJG15 r	System-level GJG15 ρ	System-level SEEDA-S r	System-level SEEDA-S ρ	System-level SEEDA-E r	System-level SEEDA-E ρ	Sentence-level GJG15 Acc	Sentence-level GJG15 τ	Sentence-level SEEDA-S Acc	Sentence-level SEEDA-S τ	Sentence-level SEEDA-E Acc	Sentence-level SEEDA-E τ
EBM	M^2	0.721	0.706	0.658	0.487	0.791	0.764	0.506	0.350	0.512	0.200	0.582	**0.328**
	Sent-M^2	0.852	0.762	0.802	0.692	0.887	0.846	0.506	0.350	0.512	0.200	0.582	**0.328**
	PT-M^2	0.912	0.853	0.845	0.769	0.896	0.909	**0.512**	0.354	**0.527**	**0.204**	**0.587**	0.293
	ERRANT	0.738	0.699	0.557	0.406	0.697	0.671	0.504	**0.356**	0.498	0.189	0.573	0.310
	SentERRANT	0.850	0.741	0.758	0.643	0.860	0.825	0.504	**0.356**	0.498	0.189	0.573	0.310
	PT-ERRANT	**0.917**	**0.886**	0.818	0.720	0.888	0.888	0.493	0.343	0.497	0.158	0.553	0.246
	GoToScorer	0.691	0.685	**0.929**	**0.881**	**0.901**	**0.937**	0.336	0.237	0.477	-0.046	0.521	0.042
SBM	GLEU	0.653	0.510	0.847	**0.886**	**0.911**	0.897	0.684	0.378	0.673	0.351	0.695	0.404
	Scribendi Score	0.890	0.923	0.631	0.641	0.830	0.848	0.498	0.009	0.354	-0.238	0.377	-0.196
	SOME	**0.975**	**0.979**	0.892	0.867	0.901	**0.951**	**0.776**	**0.555**	**0.768**	**0.555**	**0.747**	**0.512**
	IMPARA	0.961	0.965	**0.911**	0.874	0.889	0.944	0.744	0.491	0.761	0.540	0.742	0.502

Previous metrics fail to assess high-quality corrections produced by neural systems

	Metric	System-level GJG15 r	System-level GJG15 ρ	SEEDA-S r	SEEDA-S ρ	SEEDA-E r	SEEDA-E ρ	Sentence-level GJG15 Acc	GJG15 τ	SEEDA-S Acc	SEEDA-S τ	SEEDA-E Acc	SEEDA-E τ
EBM	M^2	0.721	0.706	0.658	0.487	0.791	0.764	0.506	0.350	0.512	0.200	0.582	**0.328**
	Sent-M^2	0.852	0.762	0.802	0.692	0.887	0.846	0.506	0.350	0.512	0.200	0.582	**0.328**
	PT-M^2	0.912	0.853	0.845	0.769	0.896	0.909	**0.512**	0.354	**0.527**	**0.204**	**0.587**	0.293
	ERRANT	0.738	0.699	0.557	0.406	0.697	0.671	0.504	**0.356**	0.498	0.189	0.573	0.310
	SentERRANT	0.850	0.741	0.758	0.643	0.860	0.825	0.504	**0.356**	0.498	0.189	0.573	0.310
	PT-ERRANT	**0.917**	**0.886**	0.818	0.720	0.888	0.888	0.493	0.343	0.497	0.158	0.553	0.246
	GoToScorer	0.691	0.685	**0.929**	**0.881**	**0.901**	**0.937**	0.336	0.237	0.477	-0.046	0.521	0.042
SBM	GLEU	0.653	0.510	0.847	**0.886**	**0.911**	0.897	0.684	0.378	0.673	0.351	0.695	0.404
	Scribendi Score	0.890	0.923	0.631	0.641	0.830	0.848	0.498	0.009	0.354	-0.238	0.377	-0.196
	SOME	**0.975**	**0.979**	0.892	0.867	0.901	**0.951**	**0.776**	**0.555**	**0.768**	**0.555**	**0.747**	**0.512**
	IMPARA	0.961	0.965	**0.911**	0.874	0.889	0.944	0.744	0.491	0.761	0.540	0.742	0.502

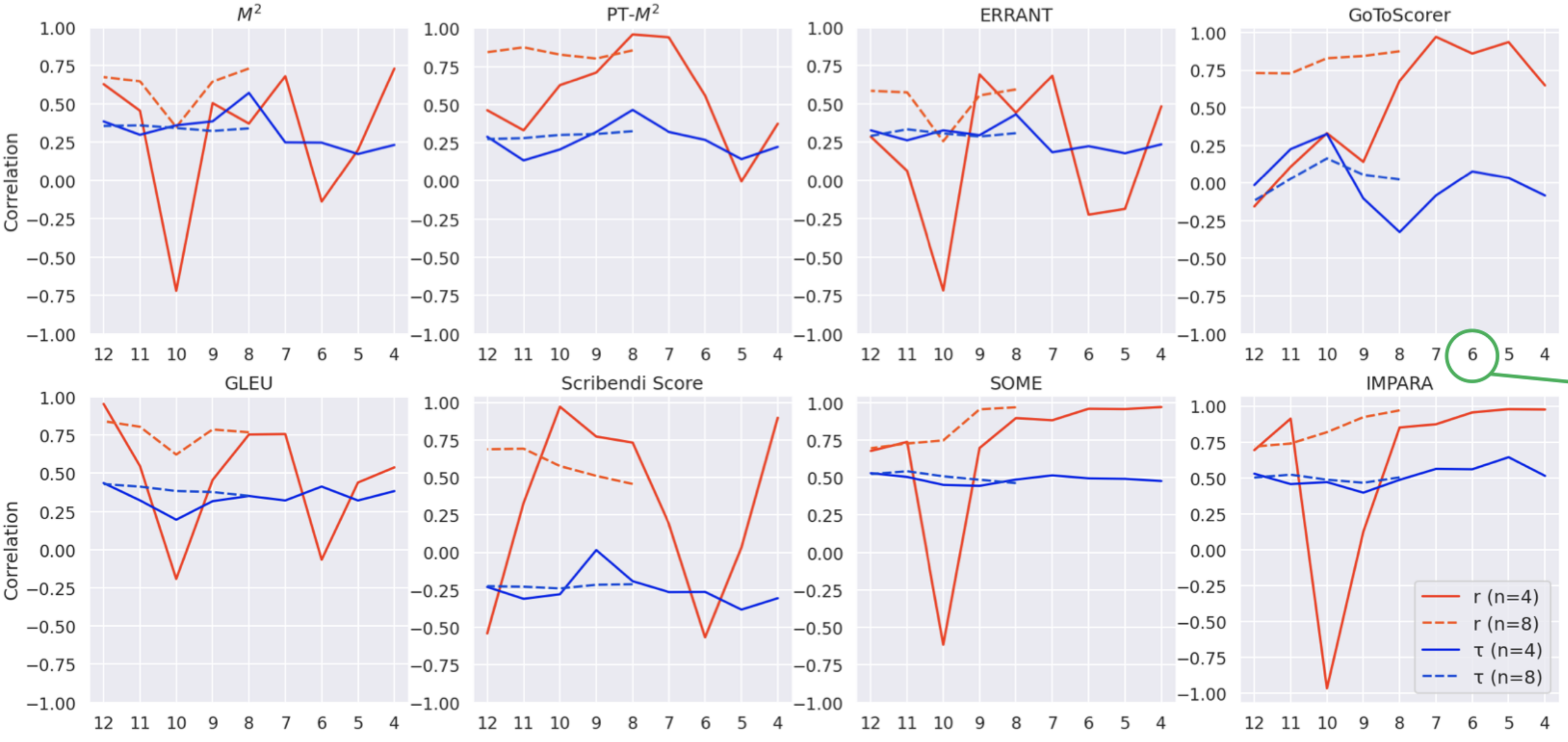
Outlier output greatly affects the meta-evaluation results

Unedited texts (+INPUT) increase correlation

fluent corrections (+REF-F, GPT-3.5) decrease correlation, esp. at the system-level meta-evaluation

Metric	System-level +INPUT r	+INPUT ρ	+REF-F, GPT-3.5 r	+REF-F, GPT-3.5 ρ	All systems r	All systems ρ	Sentence-level +INPUT Acc	+INPUT τ	+REF-F, GPT-3.5 Acc	+REF-F, GPT-3.5 τ	All systems Acc	All systems τ
M^2	0.928	0.814	-0.239	0.161	0.566	0.318	0.605	0.361	0.527	0.216	0.558	0.264
M^2 (+Min)	0.929	0.884	-0.172	0.264	0.587	0.403	0.673	0.461	0.594	0.304	0.630	0.363
M^2 (+Min, Flu)	0.930	0.880	-0.149	0.262	0.594	0.400	0.674	0.458	0.595	0.305	0.631	0.364
Sent-M^2	0.971	0.879	-0.062	0.358	0.542	0.479	0.605	0.361	0.527	0.216	0.558	0.264
PT-M^2	0.974	0.929	-0.083	0.442	0.509	0.546	0.608	0.332	0.542	0.200	0.571	0.250
ERRANT	0.925	0.742	-0.502	0.051	0.404	0.229	0.597	0.344	0.511	0.188	0.542	0.236
ERRANT (+Min)	0.922	0.753	-0.462	0.112	0.475	0.279	0.609	0.350	0.530	0.184	0.550	0.218
ERRANT (+Min, Flu)	0.920	0.725	-0.460	0.090	0.484	0.261	0.605	0.348	0.523	0.175	0.541	0.207
SentERRANT	0.965	0.863	-0.357	0.200	0.354	0.350	0.597	0.344	0.511	0.188	0.542	0.236
PT-ERRANT	0.972	0.912	-0.324	0.240	0.352	0.382	0.580	0.292	0.500	0.144	0.532	0.199
GoToScorer	0.974	0.951	0.667	0.916	0.817	0.932	0.468	-0.064	0.505	0.009	0.476	-0.048
GLEU	0.957	0.911	-0.039	0.475	0.453	0.574	0.698	0.400	0.611	0.227	0.639	0.285
GLEU (+Min)	0.868	0.942	0.236	0.704	0.593	0.760	0.758	0.519	0.662	0.327	0.685	0.372
GLEU (+Min, Flu)	0.857	0.935	0.275	0.700	0.610	0.756	0.756	0.513	0.727	0.463	0.684	0.370
Scribendi Score	0.902	0.718	0.611	0.717	0.755	0.770	0.316	-0.323	0.345	-0.264	0.315	-0.328
SOME	0.965	0.896	0.931	0.916	0.947	0.932	0.792	0.601	0.760	0.531	0.766	0.537
IMPARA	0.975	0.901	0.932	0.921	0.934	0.936	0.785	0.587	0.742	0.496	0.745	0.495

Many previous metrics fail to distinguish the performance of neural systems



#	Score	Range	System
1	0.679	1	REF-F
2	0.583	2	GPT-3.5
3	0.173	3	TransGEC
4	0.097	4-6	T5
	0.078	4-7	REF-M
	0.067	4-7	Riken-Tohoku
	0.064	4-7	BERT-fuse
5	-0.076	8-11	UEDIN-MS
	-0.084	8-11	PIE
	-0.092	8-11	GECToR-BERT
	-0.097	8-11	LM-Critic
6	-0.154	12-12	GECToR-ens
7	-0.211	13-14	TemplateGEC
	-0.231	13-14	BART
8	-0.797	15	INPUT

Human ranking

Solid lines: 4 systems, dashed lines: 8 systems, red: pearson, blue: spearman

Takeaway messages

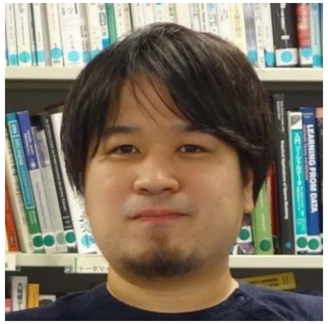
1. Edit-based models seem to be underestimated, and aligning evaluation granularity between human judgment and system output improves correlation

2. Traditional GEC evaluation metrics are not good at evaluating modern neural systems

3. Meta-evaluation should be performed thoroughly with various kinds of settings

Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction (Workshop on Innovative Use o NLP for Building Educational Applications 2024)

Joint work with Masamune Kobayashi and Masato Mita



Background

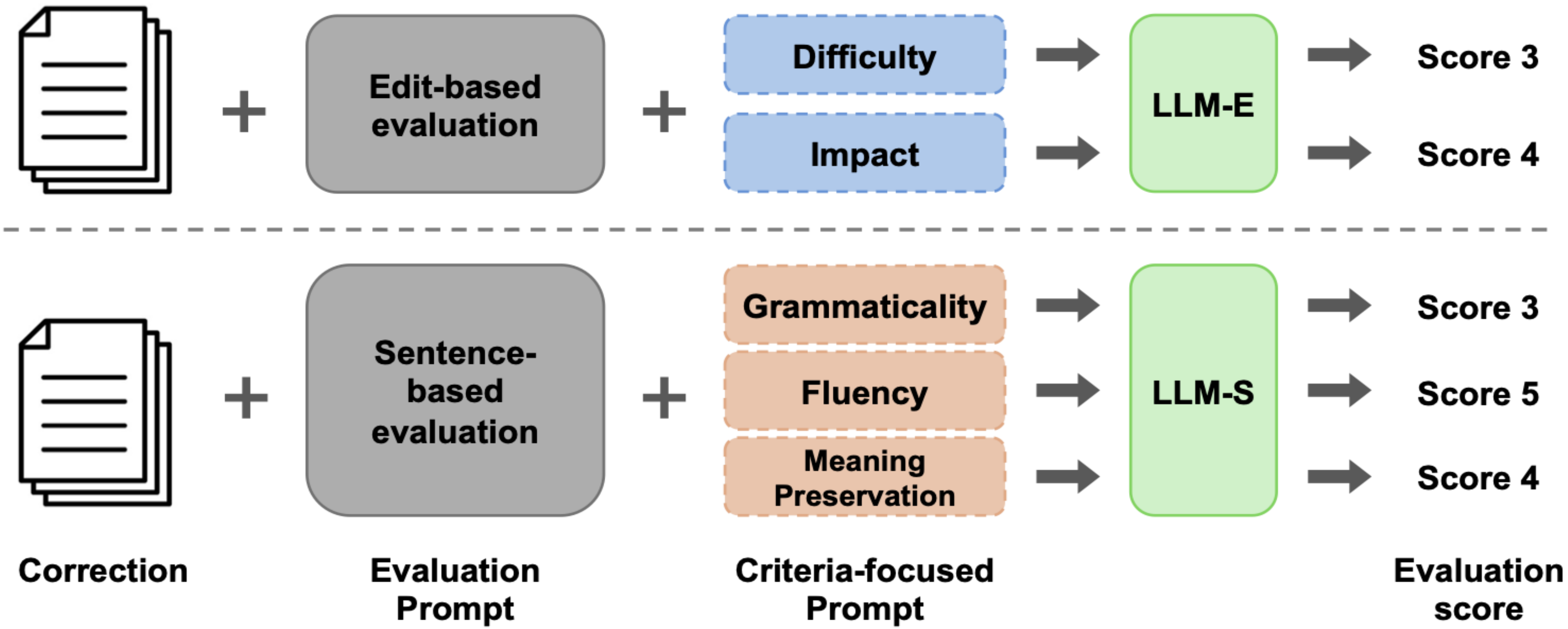
- LLMs outperform existing evaluation metrics in some tasks, such as summarization and translation
- In GEC, extensive analysis is lacking, and it is unclear how well it performs compared to existing metrics

Main findings of this work

- GPT-4 has SOTA performance compared to existing metrics
- Considering evaluation criteria in prompts leads to performance improvement (especially sentence fluency)
- As the scale of the LLMs decreased, the correlation with human evaluation decreased, and the ability to capture the fluency of corrected sentences decreased as well

Methods: LLM-as-a-judge for GEC

- LLMs evaluate the correction using prompts for each granularity focusing on evaluation criteria for GEC



Experimental setup

GEC metrics:

- Edit-based: M2, ERRANT, GoToScorer, PT-M2
- Sentence-based: GLEU, Scribendi Score, SOME, IMPARA

LLMs:

- LLaMa 2 (13B), GPT-3.5, GPT-4

Dataset: SEEDA [Kobayashi+, '24]

- Human scores are assigned at each granularity to 15 sets of sentences
- “Base” meta-evaluation: 12 outputs excluding outliers
- “+ Fluent corr.” meta-evaluation : “Base” + Two fluent corrections

Results: system-level analysis

- GPT-4 achieves the highest correlations, and criteria-focused prompts are effective.
- The correlation decreased as the LLM scale was reduced (especially in "+ Fluent corr.")
- Most of the correlations for GPT-4 exceed 0.9.

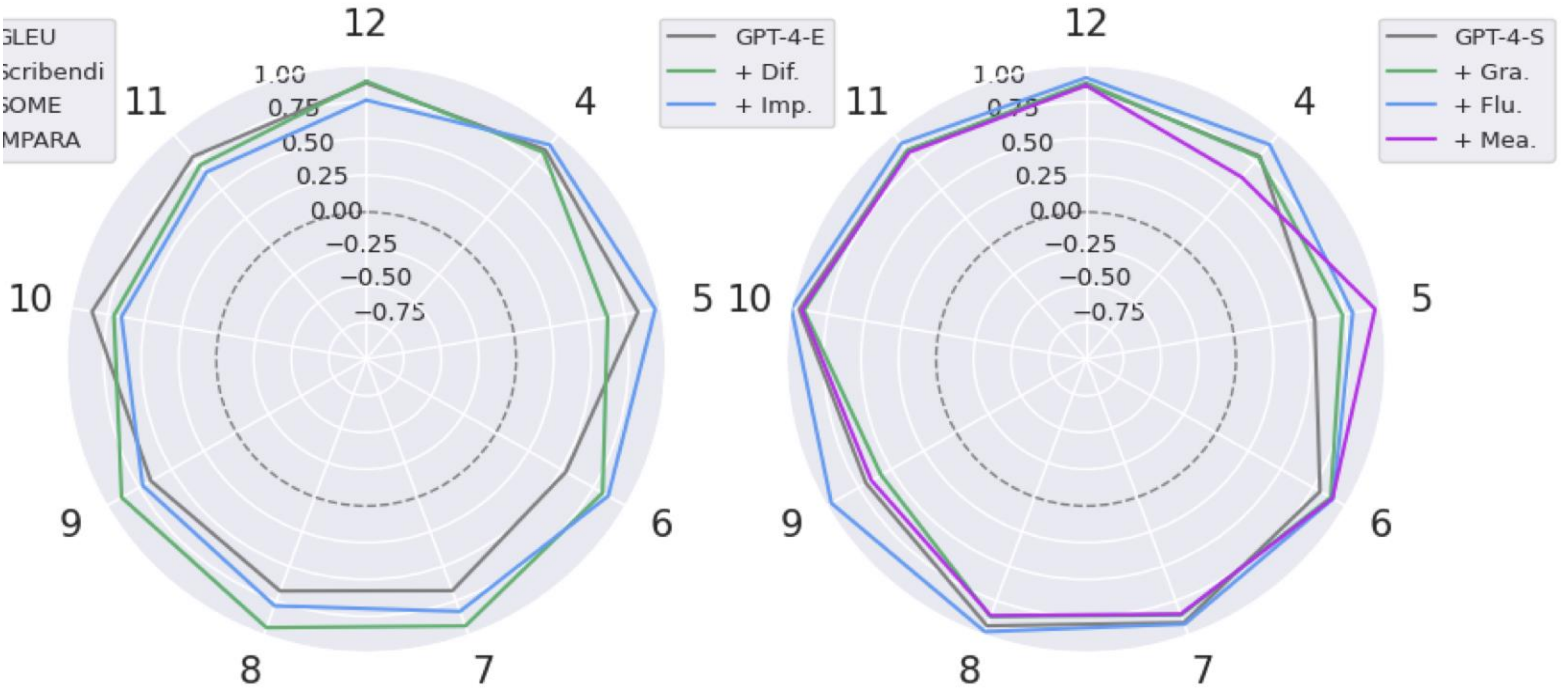
Metric	System-level							
	SEEDA-E				SEEDA-S			
	Base		+ Fluent corr.		Base		+ Fluent corr.	
	r	ρ	r	ρ	r	ρ	r	ρ
M^2	0.791	0.764	-0.239	0.161	0.658	0.487	-0.336	-0.013
ERRANT	0.697	0.671	-0.502	0.051	0.557	0.406	-0.587	-0.116
GoToScorer	0.901	0.937	0.667	0.916	0.929	0.881	0.627	0.881
PT-M^2	0.896	0.909	-0.083	0.442	0.845	0.769	-0.162	0.336
GLEU	0.911	0.897	0.053	0.482	0.847	0.886	-0.039	0.475
Scribendi Score	0.830	0.848	0.721	0.847	0.631	0.641	0.611	0.717
SOME	0.901	0.951	0.943	0.969	0.892	0.867	0.931	0.916
IMPARA	0.889	0.944	0.935	0.965	0.911	0.874	0.932	0.921
GPT-3.5-E	-0.059	0.182	-0.844	-0.257	-0.270	-0.245	-0.900	-0.525
GPT-4-E	0.911	0.965	0.845	0.974	0.839	0.846	0.786	0.899
+ Difficulty	0.941	0.972	0.909	0.978	0.885	0.860	0.863	0.908
+ Impact	0.905	**0.986**	0.848	**0.987**	0.844	0.860	0.793	0.908
Llama 2-S	0.534	0.427	0.161	0.349	0.482	0.273	0.090	0.235
GPT-3.5-S	0.878	0.916	0.302	0.648	0.770	0.636	0.199	0.433
GPT-4-S	0.960	0.958	0.967	0.969	0.887	0.860	0.931	0.908
+ Grammaticality	0.961	0.937	0.981	0.956	0.888	0.867	**0.953**	0.912
+ Fluency	**0.974**	0.979	**0.981**	0.982	0.913	0.874	0.952	0.916
+ Meaning Preservation	0.911	0.960	0.976	0.974	**0.958**	**0.881**	0.952	**0.925**

Results: sentence-level analysis

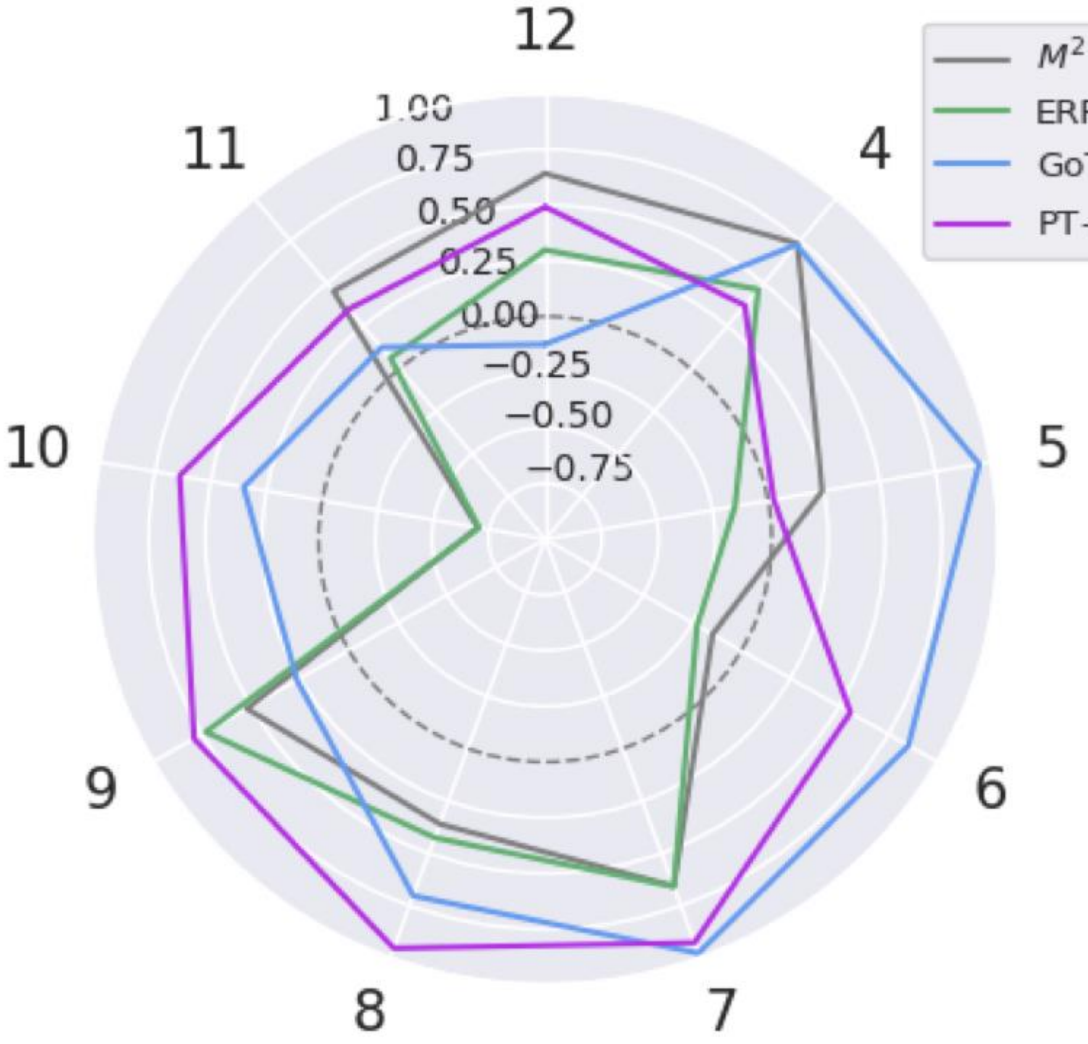
- GPT-4 performance differs from that of the system-level meta-evaluation
- “GPT-4-S + Fluency” surpassed existing metrics and achieved SOTA performance.

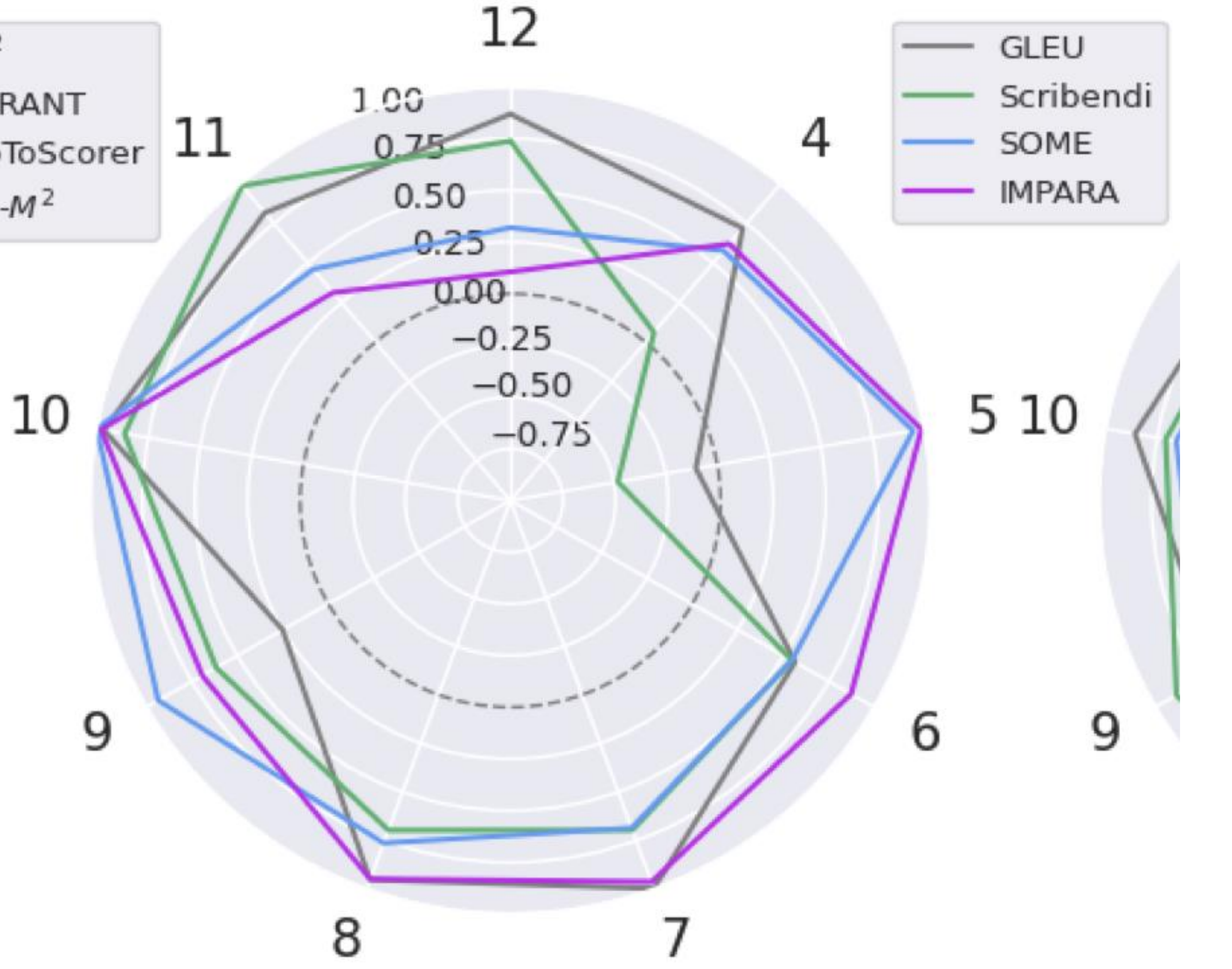
Metric	Sentence-level							
	SEEDA-E				SEEDA-S			
	Base		+ Fluent corr.		Base		+ Fluent corr.	
	Acc	τ	Acc	τ	Acc	τ	Acc	τ
M^2	0.582	0.328	0.527	0.216	0.512	0.200	0.496	0.170
ERRANT	0.573	0.310	0.511	0.188	0.498	0.189	0.471	0.129
GoToScorer	0.521	0.042	0.505	0.009	0.477	-0.046	0.504	0.009
PT-M^2	0.587	0.293	0.542	0.200	0.527	0.204	0.528	0.180
GLEU	0.695	0.404	0.630	0.266	0.673	0.351	0.611	0.227
Scribendi Score	0.377	-0.196	0.359	-0.240	0.354	-0.238	0.345	-0.264
SOME	0.747	0.512	0.743	0.494	0.768	0.555	0.760	0.531
IMPARA	0.742	0.502	0.725	0.455	0.761	0.540	0.742	0.496
GPT-3.5-E	0.463	-0.073	0.428	-0.143	0.487	-0.026	0.437	-0.126
GPT-4-E	0.728	0.455	0.702	0.404	0.698	0.395	0.687	0.374
+ Difficulty	0.719	0.437	0.708	0.417	0.717	0.434	0.703	0.406
+ Impact	0.730	0.460	0.710	0.420	0.717	0.434	0.696	0.392
Llama 2-S	0.521	0.042	0.527	0.054	0.534	0.068	0.526	0.052
GPT-3.5-S	0.633	0.265	0.597	0.195	0.631	0.263	0.608	0.216
GPT-4-S	0.798	0.595	0.783	0.565	0.784	0.567	0.770	0.540
+ Grammaticality	0.807	0.615	0.804	0.607	0.796	0.592	0.788	0.577
+ Fluency	**0.831**	**0.662**	**0.812**	**0.624**	**0.819**	**0.637**	**0.797**	**0.594**
+ Meaning Preservation	0.813	0.626	0.793	0.587	0.810	0.620	0.792	0.584

System-level window analysis of higher-ranking systems: GPT-4-S works best



System-level window: conventional metrics are not robust for neural GEC models





Two research questions for GEC evaluation

Are the existing datasets adequate in the era of deep learning?

→Revisiting meta-evaluation (evaluation of evaluations) for grammatical error correction

Can LLMs be used to evaluate grammatical error correction?

→Application of LLMs for evaluation of grammatical error correction (LLM-as-a-judge)

References

- Masamune Kobayashi, Masato Mita, Mamoru Komachi. **Revisiting Meta-evaluation for Grammatical Error Correction**. (TACL 2024) PDF
- Masamune Kobayashi, Masato Mita, Mamoru Komachi. **Large Language Models are State-of-the-Art Evaluator for Grammatical Error Correction.** (BEA 2024) PDF