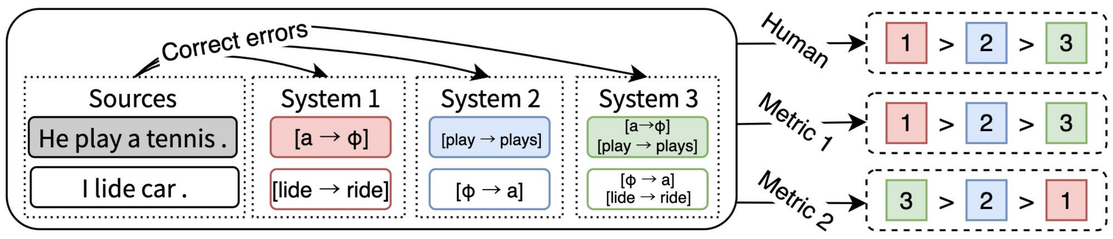
Rethinking Evaluation Metrics for Grammatical Error Correction: Why Use a Different Evaluation Process than Human?

Takumi Goto, Yusuke Sakai, Taro Watanabe *NARA Institute of Science and Technology*

Automatic Grammatical Error Correction (GEC) Evaluation

- One Role: System ranking
  - e.g., Select the best system from a system set
- The ranking that closes human eval. is desirable
  - Meta-evaluation compares to human eval.

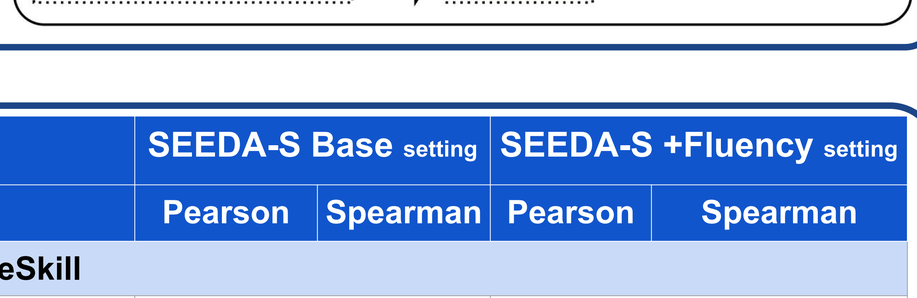


Problem of Current Automatic Evaluation

Gap of aggregation between automatic and human

- Human uses relative evaluation:
  A rating algorithm (e.g., TrueSkill) aggregates from pairwise comparison results
- Automatic one uses absolute evaluation:
  Aggregate sentence-level scores by averaging sentence-level scores. Then, sort them to obtain rankings

Why automatic evaluation does not use the same aggregation with human's?



Experiments and Results

- **Resolve the gap by performing automatic evaluations using the same method as human evaluation**
- Metrics

Edit-level:	ERRANT, PT-ERRANT
n-gram-level:	GLEU, GREEN
Sent-level:	SOME, IMPARA, Scribendi(omitted)

- Meta-evaluation: SEEDA dataset [Kobayashi+24]
  - Includes 14 GEC systems and human's rankings
  - Base setting: Exclude some system, e.g., GPT-3.5
  - +Fluency setting: Use all of 14 systems
  - Metrics: Pearson and Spearman correlations
- Resolving gap improves correlations
  - E.g., IMPARA outperforms GPT-4 in +Fluency, suggesting existing metrics have been underestimated.
  - There is no effect on n-gram metrics, because their reliability of the sentence-level scores

	SEEDA-S Base setting		SEEDA-S +Fluency setting	
Metrics	**Pearson**	**Spearman**	**Pearson**	**Spearman**
w/o TrueSkill				
ERRANT	0.545	0.343	-0.591	-0.156
PT-ERRANT	0.700	0.629	-0.546	0.077
GLEU	0.886	0.902	0.155	0.543
GREEN	0.925	0.881	0.185	0.569
SOME	0.892	0.867	0.931	0.916
IMPARA	0.916	0.902	0.887	0.938
w/ TrueSkill (Proposal)				
ERRANT	0.763	0.706	-0.463	0.095
PT-ERRANT	0.870	0.797	-0.366	0.182
GLEU	0.863	0.846	0.017	0.393
GREEN	0.855	0.846	-0.214	0.327
SOME	0.932	0.881	0.971	0.925
IMPARA	0.939	**0.923**	**0.975**	**0.952**
LLM-based metrics [Kobayashi+ 24]				
GPT-4-E (flu.)	0.844	0.860	0.793	0.908
GPT-4-S (flu.)	0.913	0.874	0.952	0.916
GPT-4-S (mea.)	**0.958**	0.881	0.952	0.925

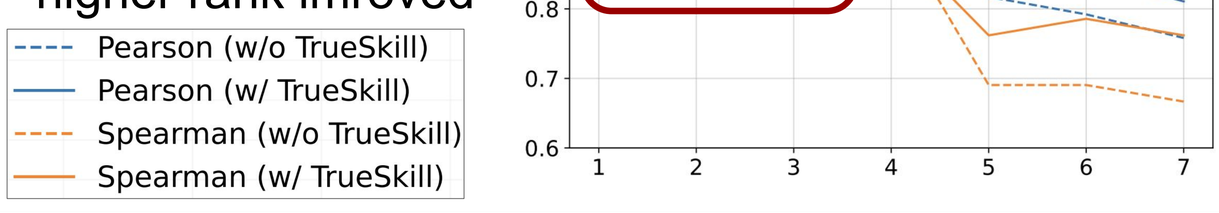
[Kobayashi+ 24]: Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction,

Analysis / Discussion

How does resolving gap improve evaluation?

- We use window-analysis proposed in SEEDA
- Correlation from Xth-rank to {X+7}th-rank in human evaluation

- Correlation for higher rank imroved



Suggestion of developing and using metrics

- **Use:** Employ the same aggregation as human evaluation for ranking systems
  - Currently human ranking uses TrueSkill, so metrics should also use TrueSkill
- **Develop:** Metrics can provide high-quality sentence-level scores are needed
  - Score quality directly affect ranking quality