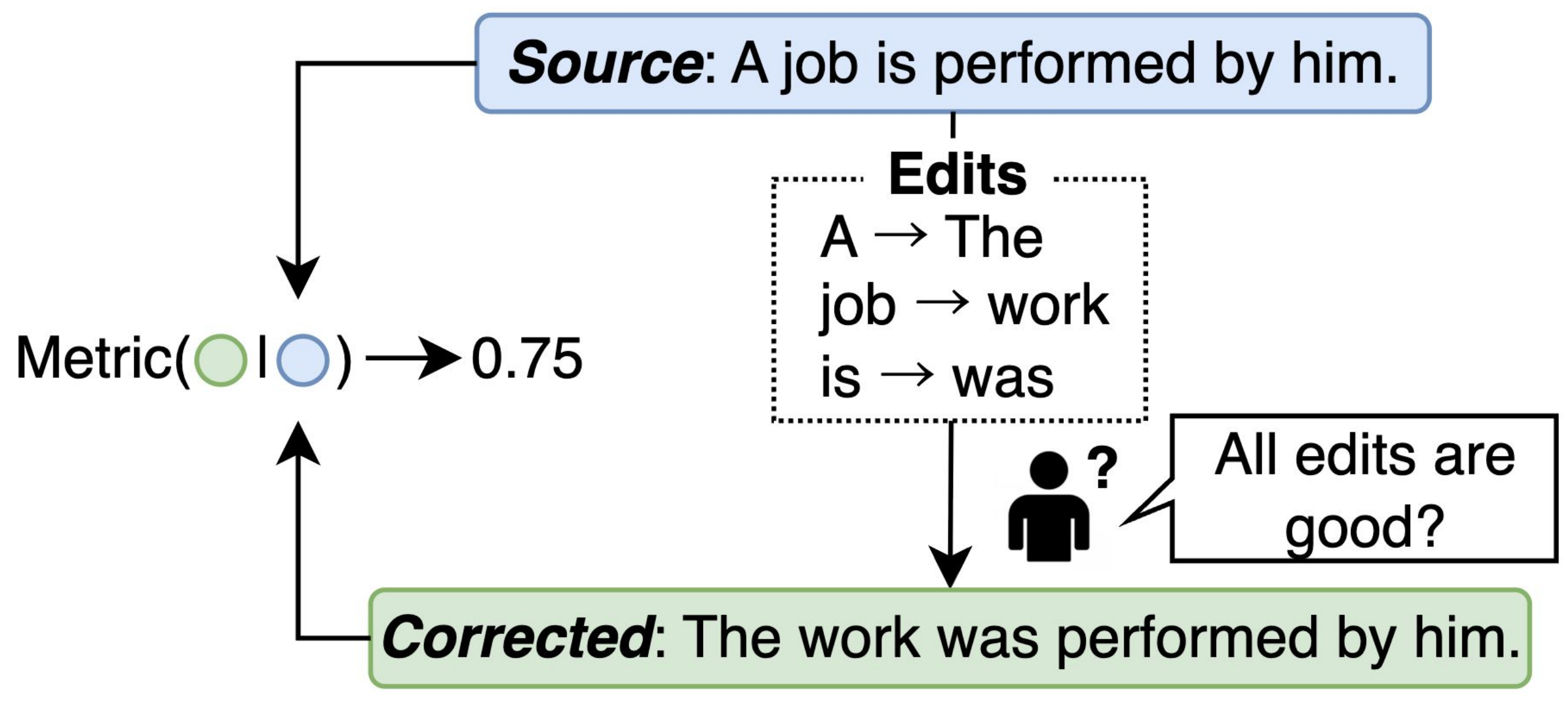
Improving Explainability of Sentence-level Metrics via Edit-level Attribution for Grammatical Error Correction

Takumi Goto, Justin Vasselli, Taro Watanabe *NARA Institute of Science and Technology*

Background

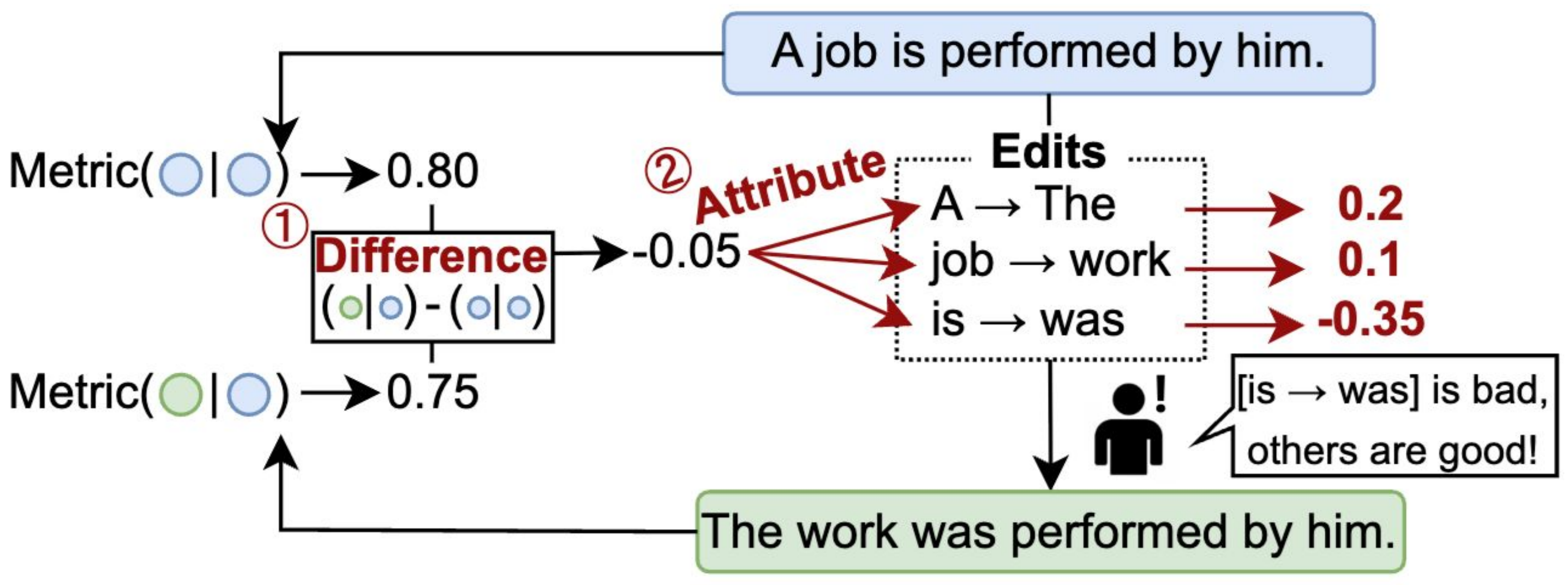
- Recent sentence-level metrics achieve higher evaluation performance
  - ○ E.g., SOME[Yoshimura, COLING2020], IMPARA[Maeda+ COLING2020]
- However, they provide only a single scalar as a score, thus their **explainability is low**
- Researchers cannot use such metrics for analysis purpose, and users cannot receive detailed feedback


Source: A job is performed by him.
Edits
A → The
job → work
is → was
Metric(|) → 0.75
?
All edits are good?
Corrected: The work was performed by him.

Proposed method: Edit-level attribution

- Feature attribution is a fundamental approach to improve explanability
- We regard edits as features, and **attribute sentence-level score to edits**

- Attribution steps
  - ○ **1. Quantify the entire contributions of edits** by taking difference: $\Delta M(H|S) = \underbrace{M(H|S)}_{\text{Hypothesis score}} - \underbrace{M(S|S)}_{\text{Source socre}}$
  - ○ **2. Attribute it to the edits via Shapley values** For i-th edit, consider all apply patterns of other than i-th edit, and quantify its contribution



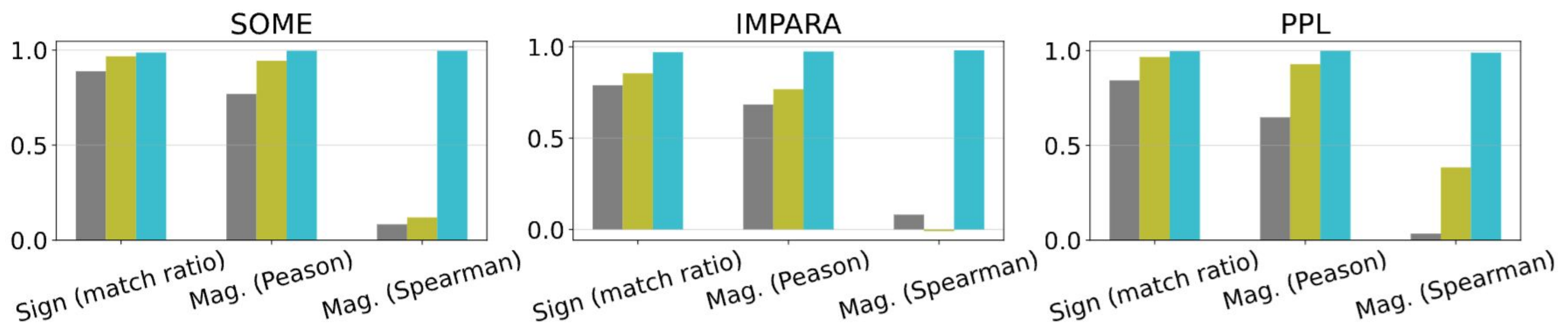
$$\phi_i(M) = \sum_{\boldsymbol{e}' \subseteq \boldsymbol{e} \setminus \{e_i\}} \frac{|\boldsymbol{e}'|!(N - |\boldsymbol{e}'| - 1)!}{N!} (\underbrace{\Delta M(S_{\boldsymbol{e}' \cup \{e_i\}}|S)}_{\text{Score with } \boldsymbol{e}} - \underbrace{\Delta M(S_{\boldsymbol{e}'}|S)}_{\text{Score without } \boldsymbol{e}})$$

- The attribution scores provide:
  - ○ **Sign**: Which edit improves or worsens the sentence-level score?
  - ○ **Magnitude**: How much improvement or deterioration?

Evaluation of attribution scores

Faithfulness: How well the attribution results reflect the model's internal decision

- Compare individual edits and grouped edits



(b) JFLEG results.

Explainability: The extent to which the results are understandable to humans

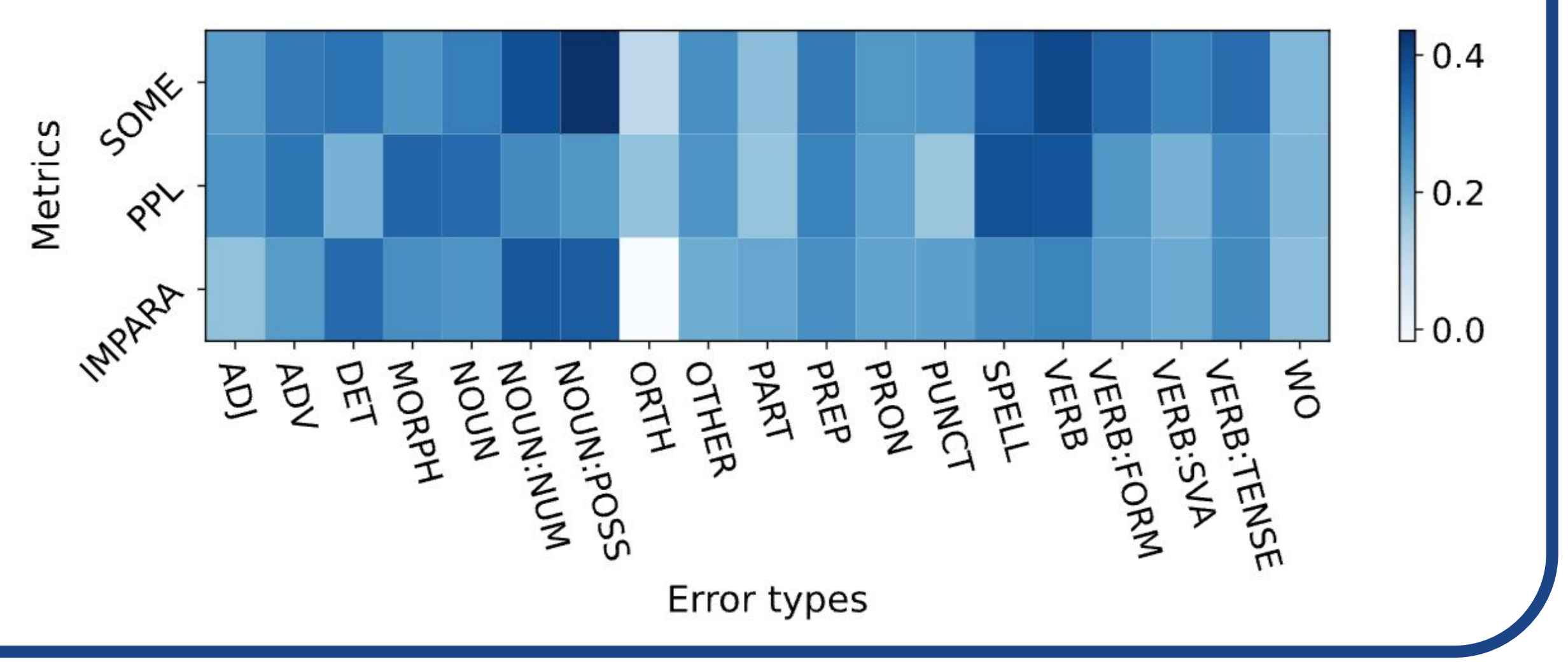
- Compare pos/neg of scores and correct/incorrect of human evaluation
  - ○ Use SEEDA as a human evaluation
- Confirm 60%~70% agreement
  - ○ (Random baseline is 50%)

Applications of the attributed results

- **Case study** explains a sentence-level scores
  - ○ E.g., Why IMPARA's score worsens? → the edit [u → you] is the reason.

Original (S)	-	Further more		by	these	evidence		u	will agree	
Correction (H)	-	Further more	,	with	this	evidence	,	you	will agree	.
Metrics (M)	$\Delta M(\cdot)$	Shapley values $\phi_i(M)$								
SOME	0.298	-	0.068	0.064	0.033	-	0.038	0.066	-	0.030
IMPARA	-0.027	-	0.068	0.029	0.124	-	0.145	-0.361	-	-0.033
PPL	1266.3	-	250.7	103.8	216.0	-	67.4	366.6	-	261.5

- **Reveal the bias** of metric: What kind of edits do metrics prefer?
  - ○ Compute average attribution scores for Error Type[Bryant+ ACL2017]
  - ○ E.g., ORTH (Orthography, case or whitespace) is underrated


Metrics
SOME
PPL
IMPARA
ADJ
ADV
DET
MORPH
NOUN
NOUN:NUM
NOUN:POSS
ORTH
OTHER
PART
PREP
PRON
PUNCT
SPELL
VERB
VERB:FORM
VERB:SVA
VERB:TENSE
WO
Error types