P3-9

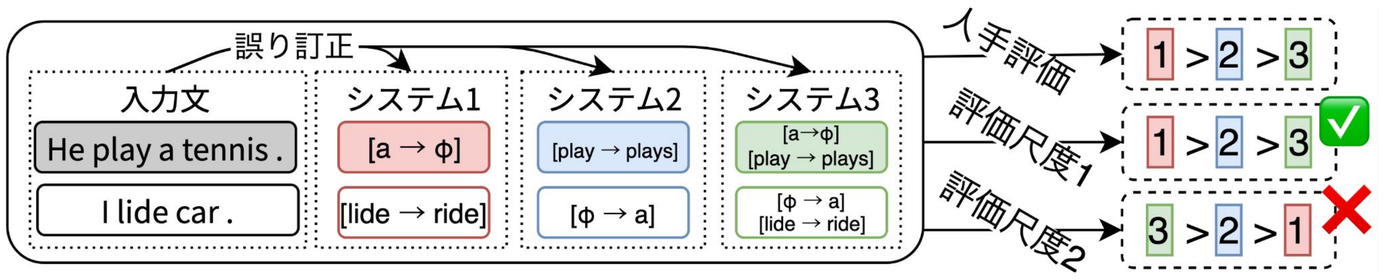
文法誤り訂正における人手評価と自動評価の乖離とその解決

五藤巧，坂井 優介，渡辺太郎　*奈良先端科学技術大学院大学*

コードも公開!

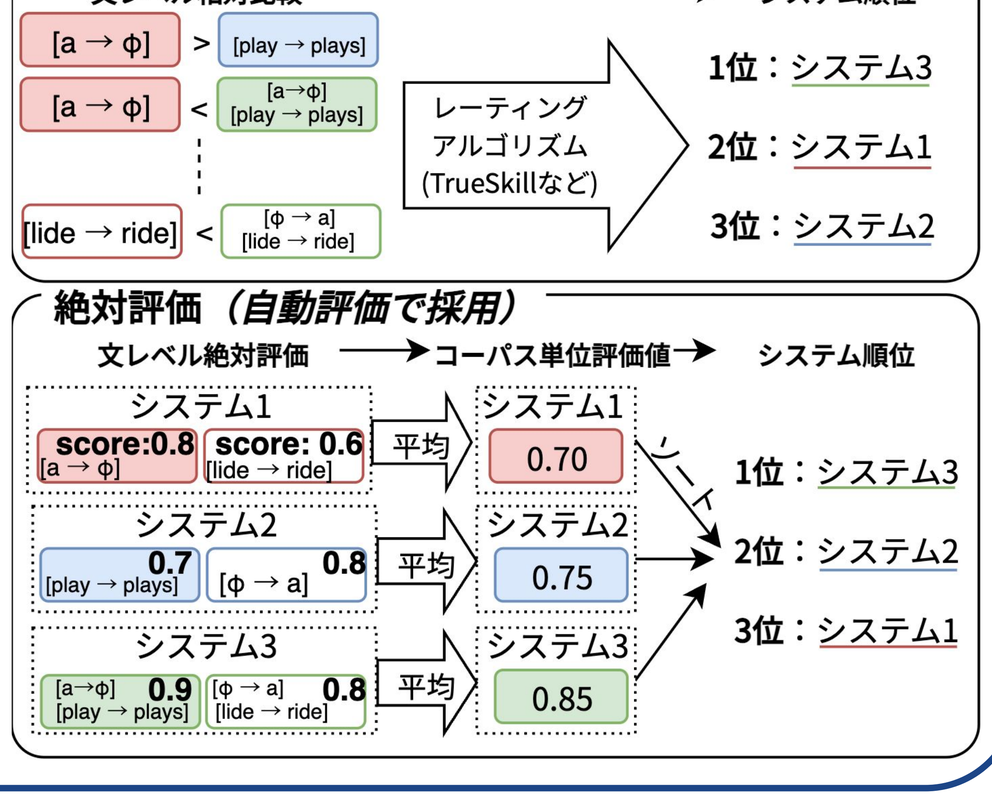
文法誤り訂正における評価

- **一つの目的**：訂正システムの順位付け
  - 大量のシステムからよいものを選ぶ
- **人手評価を模倣した順位が望ましい**
  - 実際に人手評価との相関でメタ評価



現状の自動評価の問題とその解決

- **問題：自動評価の順位の計算方法が人手評価と乖離**
- 人手評価は相対評価：
  文単位の比較結果から
  レーティングアルゴリズム (TrueSkillなど) で順位を計算
- 自動評価は絶対評価：
  文単位の絶対評価値から
  コーパス単位の評価値を経由して順位を計算
- 自動評価は人手評価の模倣を目的としているのに，
  なぜ順位の計算方法は模倣していないのか？
  →人手評価と同じ方法で順位を計算することで解決



実験と結果

乖離を埋めることで人手評価との相関が向上するか？

- 評価尺度
  - 編集レベル尺度：ERRANT，PT-ERRANT
  - n-gramレベル尺度：GLEU，GREEN
  - 文レベル尺度：SOME，IMPARA，Scribendi
- メタ評価：SEEDAデータセット
  - 多様なモデルを含む14システムの出力と
    人手評価の順位を含むデータ
    - Base設定: GPT-3.5などを除いた12システム
    - +Fluency設定: GPT-3.5などを含む14システム
  - 評価尺度と人手評価の相関：PearsonとSpearman
- 計算方法の乖離を埋めることで相関が向上
  - 人手評価をより模倣した評価が可能に
  - +Fluency設定でIMPARAはGPT-4を超える
  - n-gramレベル尺度では効果なし
    - 人手評価はn-gramレベルではないことに起因？

	SEEDA-S Base設定		SEEDA-S +Fluency設定	
評価尺度	**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 (提案法)				
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**
大規模言語モデルによる評価 [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

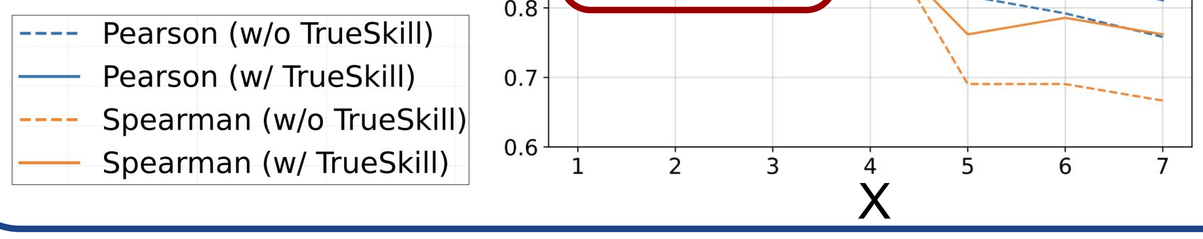
[Kobayashi+ 24]: Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction,

分析・議論

乖離の解決はどのような点を改善するか？

- SEEDAで提案されたwindow-analysis
- 人手評価のX位からX+7位に限定した相関

- 上位システムの
  Pearsonが特に改善



今後の評価尺度の使用と開発について

- **使用**：評価尺度でシステムを順位付けする時，
  人手評価と同じ方法で計算するべき
  - コーパス単位の評価値を使う従来方法は
    尺度の評価性能を最大限に引き出せない
- **開発**：文の相対評価を正確に行える尺度が必要
  - 相対評価が完璧にできればレーティング
    アルゴリズムで計算される順位も完璧に