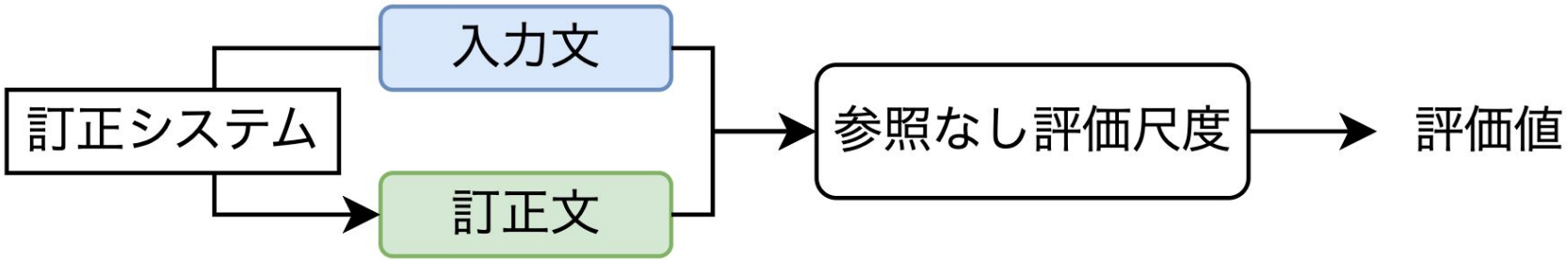
文法誤り訂正における参照なし評価尺度を用いた分析的評価法

○五藤巧, 渡辺太郎 (NAIST)

文法誤り訂正における参照なし評価

- **文法誤り訂正タスク**: This are gramatical sentence. → This is a grammatical sentence.
- **参照なし評価尺度**: 推定された訂正文を正解文を用いずに評価

  利点:
  - 正解文のアノテーションが必要なし
  - 正解の訂正を網羅的に評価可能
    - 参照あり評価では，モデルの訂正が正しくても正解文に含まれなければ誤りに

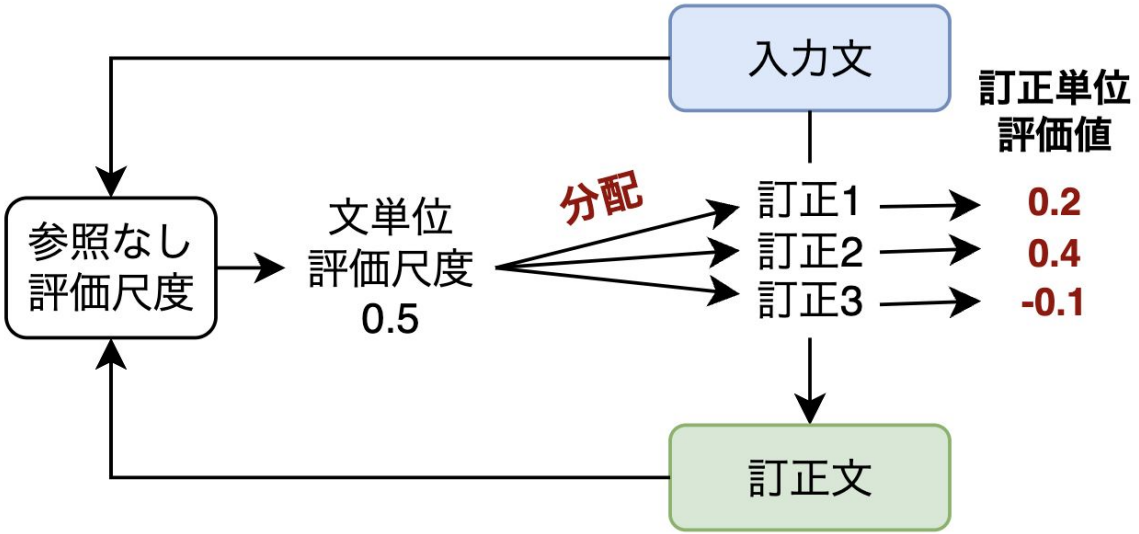


参照なし評価の問題点

- 評価値の向上・低下の要因がどの訂正にあるのか分からない
  - 参照なし評価尺度は，文単位の評価値を一つの実数で計算するのみ
  - 訂正単位の評価値は得られない
- これにより..
  - 定性的な分析に繋がらない
  - どのようにシステムを改良すればよいかの知見が得られない

提案法の概要

- 文単位の評価値を訂正単位の評価値に分配する**分析的評価法**を提案
  - 訂正単位の評価値の総和が文単位評価値に一致するように分配
  - 評価値の正負 → 訂正の良し悪し
  - 評価値の絶対値の大きさ → 良さ・悪さの度合い



シャープレイ値に基づく分配

- 分配方法にはシャープレイ値[Shapley+ 53]を応用
  - 協力ゲーム理論で用いられる
  - 全体の利益をプレイヤーの貢献度に応じて分配した時の利益

文単位の評価値　訂正　訂正単位の評価値

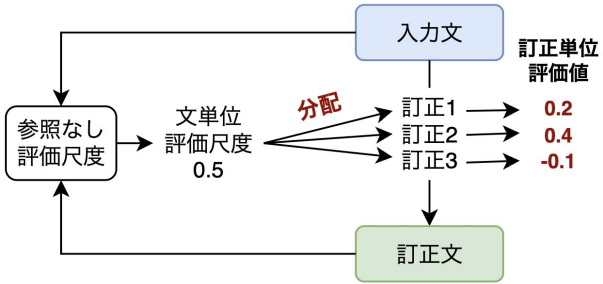
- 分配元：文単位の評価値

$$\Delta M(S,H) = M(S,H) - M(S,S)$$

尺度　入力文　訂正文

訂正文の評価値　何も訂正しない場合の評価値

- 分配先：SをHにするための各訂正



[Shapley+ 53]: Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.

シャープレイ値の計算

- ある一つの訂正を適用した時と適用しない時の評価値の差分に基づく
  - 考えうる全ての訂正集合について計算し，重みつき和を計算
  - N 個の訂正 $\boldsymbol{e} = \{e_i\}_{i=1}^{N}$ があるとき

$$\phi_i(M) = \sum_{E \subseteq \boldsymbol{e} \setminus e_i} \frac{|E|!(N - |E| - 1)!}{N!} (\Delta M(S, S_{E \cup e_i}) - \Delta M(S, S_E))$$

シャープレイ値の計算

- ある一つの訂正を適用した時と適用しない時の評価値の差分に基づく
  - 考えうる全ての訂正集合について計算し，重みつき和を計算
  - N 個の訂正 $\boldsymbol{e} = \{e_i\}_{i=1}^{N}$ があるとき

$\phi_i(M) =$

尺度Mにおける
i番目の訂正の評価値は

$$\sum_{E \subseteq \boldsymbol{e} \setminus e_i} \frac{|E|!(N - |E| - 1)!}{N!} (\Delta M(S, S_{E \cup e_i}) - \Delta M(S, S_E))$$

シャープレイ値の計算

- ある一つの訂正を適用した時と適用しない時の評価値の差分に基づく
  - 考えうる全ての訂正集合について計算し，重みつき和を計算
  - N 個の訂正 $\boldsymbol{e} = \{e_i\}_{i=1}^{N}$ があるとき

尺度Mにおける
i番目の訂正の評価値は

$$\phi_i(M) = \sum_{E \subseteq \boldsymbol{e} \setminus e_i} \frac{|E|!(N - |E| - 1)!}{N!} (\Delta M(S, S_{E \cup e_i}) - \Delta M(S, S_E))$$

i番目の訂正を除いた
訂正集合の全ての部分集合
について

シャープレイ値の計算

- ある一つの訂正を適用した時と適用しない時の評価値の差分に基づく
  - 考えうる全ての訂正集合について計算し，重みつき和を計算
  - N 個の訂正 $\boldsymbol{e} = \{e_i\}_{i=1}^{N}$ があるとき

尺度Mにおける
i番目の訂正の評価値は

$$\phi_i(M) = \sum_{E \subseteq \boldsymbol{e} \setminus e_i} \frac{|E|!(N - |E| - 1)!}{N!} \left(\Delta M(S, S_{E \cup e_i}) - \Delta M(S, S_E) \right)$$

i番目の訂正を除いた
訂正集合の全ての部分集合
について

Eにi番目の訂正を
加えて訂正したときの評価値と

加えないで訂正したとき
の評価値の

差に基づく

シャープレイ値の計算の例示

- 訂正が3つある状況を仮定：e1, e2, e3
- e1に対する評価値が知りたい場合，以下のように計算

$$\begin{aligned}\phi_1(M) = [&\frac{1}{3}(\Delta M(S, S_{\{e_1\}}) - \Delta M(S, S)) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_1,e_2\}}) - \Delta M(S, S_{\{e_2\}})) \\ &+ \frac{1}{6}(\Delta M(S, S_{\{e_1,e_3\}}) - \Delta M(S, S_{\{e_3\}})) \\ &+ \frac{1}{3}(\Delta M(S, S_{\{e_1,e_2,e_3\}}) - \Delta M(S, S_{\{e_2,e_3\}}))]\end{aligned}$$

シャープレイ値の性質と文法誤り訂正における解釈

- **効率性**
  訂正単位の評価値の総和が文単位の評価値に一致
- 対称性
  文単位評価値への貢献が同じ度合いの訂正の評価値は一致
- **ダミープレイヤ**
  評価値に貢献しない訂正の評価値はゼロ
  （「無訂正」という編集の存在を仮定してもダミープレイヤになるため，結果に影響しない）
- 加法性
  複数の観点を統合した場合に訂正の評価値を知りたいとき，
  先に評価値を統合してからシャープレイ値を計算しても，
  個別にシャープレイ値を計算してから後で統合しても結果は一致

実験

- 参照なし評価尺度
  - SOME [Yoshimura+ 20]:
    人手の評価結果に直接最適化することで評価器を構築
    - 流暢性(SOME-f)・文法性(SOME-g)・意味保存性(SOME-m)の3つの観点で評価
  - IMPARA [Maeda+ 20]:
    訂正前後文のBERT埋め込み類似度を推定できるように評価器を学習
- 入力文・訂正文のサンプルを得るためのデータセット・モデル
  - データセット：JFLEG-dev [Napoles+ 17] (754文)
  - 訂正モデル：GECToR [Omelianchuk+ 20] (RoBERTaベース)で推定

[Yoshimura+20]: SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction (Yoshimura et al., COLING 2020)
[Maeda+20]: IMPARA: Impact-Based Metric for GEC Using Parallel Data (Maeda et al., COLING 2022)
[Omelianchuk+ 20]: GECToR – Grammatical Error Correction: Tag, Not Rewrite (Omelianchuk et al., BEA 2020)
[Napoles+ 17]: JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction (Napoles et al., EACL 2017)

分析的評価の例示

- 目的：文単位評価値の向上・低下がどの訂正に起因するか知りたい
  ΔM(S, H)

- SOME-{fg}では評価値が向上：isの削除やratherの削除が主に貢献
- IMPARAは評価値が低下：only→justの訂正が悪影響で要改善
- 下図のような可視化をもって，本稿の目的は達成

表 1　実際の訂正を用いた提案法の適用例.

原文	-	It	is	also	take	risks	raher	than	only	doing		.
訂正文	-	It		also	takes	risks		than	just	doing	things	.
尺度	$\Delta M(S,H)$	編集単位の評価値 ϕ_i										
SOME-f	0.3214	-	0.090	-	0.038	-	0.101	-	0.045	-	0.048	-
SOME-g	0.2427	-	0.064	-	0.024	-	0.075	-	0.042	-	0.037	-
SOME-m	0.2332	-	0.002	-	0.008	-	0.124	-	0.005	-	0.095	-
IMPARA	-0.0004	-	0.024	-	0.066	-	0.066	-	-0.155	-	-0.002	-

提案法の妥当性

- 提案法により得られた訂正単位の評価値は妥当なのか？
- 人手評価は望ましくない
  - 人手評価を模倣することは目的としていない
  - 人手評価と一致していなくても，尺度が捉える結果を知りたい，
    というのが本稿での分析的評価の位置付け

- シャープレイ値の計算方法や性質から妥当性を主張
  - 計算方法から，周辺の訂正との依存関係を考慮して評価値を計算可能
  - 効率性の性質が，提案法の「分配したい」要求を自然に満たす

提案法のその他の解釈および展望

- **メタ評価法**
  - どのような訂正に高い・低い評価値が計算されるかを分析することで，尺度の特徴を知ることができる.
  - 従来の人手評価との相関に基づく方法と異なり，評価したい項目に敏感に反応するような尺度を選択するために役立つ.
- **バイアス・脆弱性評価法**
  - バイアス：性別や国籍などの観点で変化させたとき訂正の評価値が変わるか？
  - 脆弱性：そもそも訂正でないような編集を不当に高く評価してしまわないか？
- **説明性手法**
  - 従来手法が入力単語そのものの予測値への貢献度を説明するのに対し，提案法では「訂正」という変更操作に対して説明するところに面白さがある

まとめ

- **問題提起**:
  従来の参照なし評価尺度は文単位の評価値しか計算できず，それが向上・低下した理由を分析できない問題を指摘
- **解決法**:
  文単位評価値を訂正単位の評価値に分配する分析的評価法を提案
- **実験**:
  実例を用いて分析的評価の例を提示
- **考察など**:
  提案法の妥当性・解釈および展望について議論

付録

関連研究

- 文法誤り訂正では訂正単位の評価値(やそれに類似する値)を計算する試みが存在

手法	定式化(ざっくり)	定式化のお気持ち	提案法との差分として代表的な特徴
提案法	$S_{E \cup e_i}$ と S_E の比較（$E \subseteq \boldsymbol{e} \setminus e_i$）	e_iの有無の全通り	**手法:尺度出力の差分** **目的:尺度が考える評価値の可視化**
IMPARA のimpact計算	$S_{\boldsymbol{e}}$ と $S_{\boldsymbol{e} \setminus e_i}$ の比較	完全な訂正から e_iのみ除く	手法:埋め込み類似度 目的:人間らしい尺度開発
PT-M2	S と S_{e_i} の比較	入力文に e_iのみ加える	手法:B{EA}RTScoreの差分 目的:人間らしい尺度開発
永田ら	S と S_{e_i} の比較	入力文に e_iのみ加える	手法:埋め込み類似度 目的:訂正重要度の定量化

IMPARA: IMPARA: Impact-Based Metric for GEC Using Parallel Data (Maeda et al., COLING 2022)
PT-M2: Revisiting Grammatical Error Correction Evaluation and Beyond (Gong et al., EMNLP 2022)
永田ら: 文法誤り訂正への訂正重要度の導入(言語処理学会 第28回年次大会)

計算量

- 厳密なシャープレイ値の計算量はO(2^N)
  - Nは訂正の数
- JFLEG-devでは訂正の数は高々14であるため，現実的な時間で厳密なシャープレイ値が計算可能であった
  - より多くの訂正を含む場合，近似計算をする必要あり（future work）