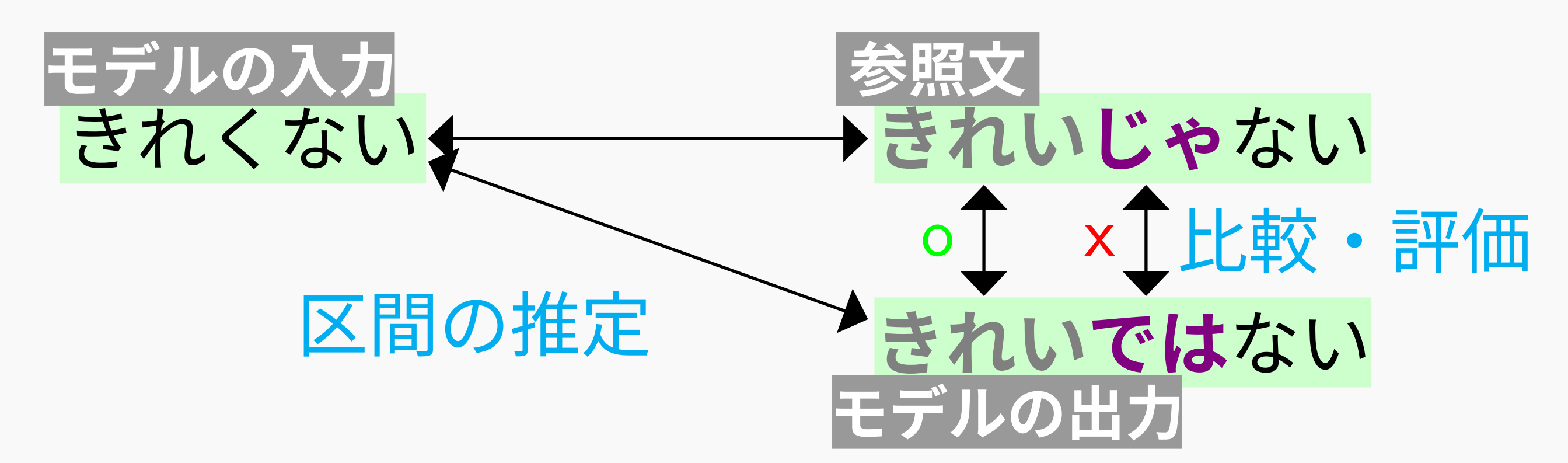


日本語誤り訂正の(評価の)ための誤り区間と誤り種類の自動アノテーションに向けて

P4-9 古山翔太^{1,2}, 永田亮³, 高村大也², 岡崎直観^{1,2}
東工大 D1 (1 東工大, 2 産総研, 3 甲南大)

①モデル化：誤り訂正の評価

- ▶ 入力
 - ▶ モデルの入力・モデルの出力・参照文
- ▶ 計算するもの
 - ▶ 誤り区間の対応関係 ←これをやる
- ▶ 出力
 - ▶ モデルの出力と参照文での一致スコア



②誤り種類はなぜ必要？

- ① 誤り訂正モデルの評価 (NLP 分野)
 - ▶ 誤り種類別の詳細な評価が可能
- ② 誤用分析 (言語教育分野)
 - ▶ 大規模学習者コーパス分析の自動化



③アノテーション基準：誤り区間

- ▶ 一貫性のある基準として、以下に限定
 - ▶ 単語・複合語・記号
 - ▶ 助詞・助動詞の接続
 - ▶ multi-word expression (MWE)
 - ▶ 文節

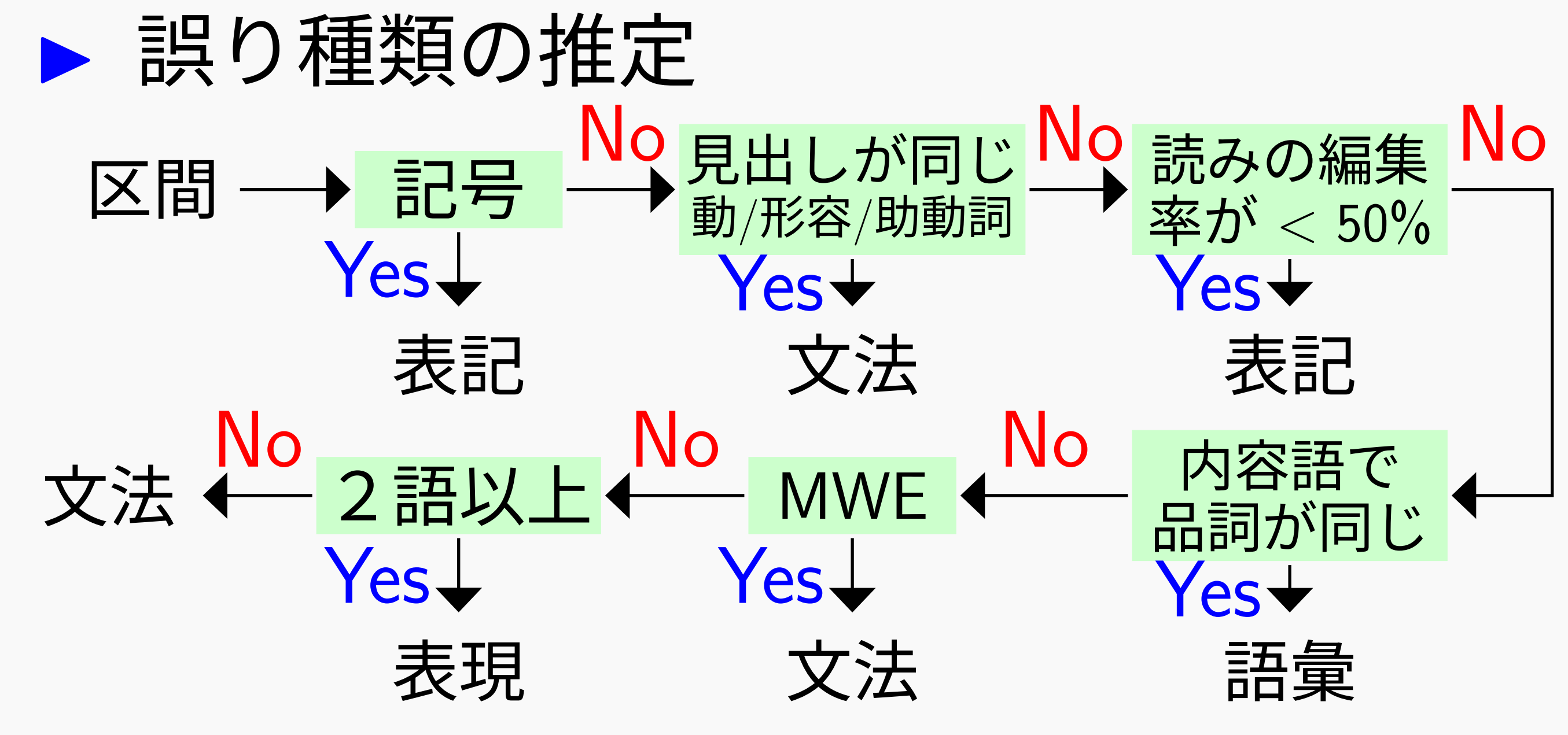
④アノテーション基準：誤り種類

- ▶ 言語学的分類 (4 種類)
 - ▶ 表記・語彙・文法・表現・その他

⑤誤り区間の推定

- ▶ ERRANTの方法をもとに手法
 - ▶ 品詞等を考慮した編集距離と動的計画法で単語間の編集を得て，編集をマージして誤り区間を得る
- ▶ 編集距離の設計 (単語分割は spaCy/GiNZA)
$$\text{距離} = \begin{cases} 0 & (\text{同じトークン}) \\ 0.5 & (\text{削除 or 挿入}) \\ 0.333 \times (\text{読みの編集率} + \text{見出し} + \text{品詞}) & \end{cases}$$
 - ▶ 「読み」が英語との差分
- ▶ マージ規則
 - マージ → 隣接する編集が Yes → MWE か → する
 - No → しない

⑥誤り種類の推定



⑦データ作成・評価

- ▶ TEC-JL コーパス中の 100 文に対して，アノテーションし，評価
 - ▶ 改善の余地が大きく，課題が残る

	TP	FP	FN	Prec	Rec	F ₁
区間	174	108	71	61.70	71.02	66.03
区間＋種類	TP	FP	FN	Prec	Rec	F ₁
表記	21	9	27	70.00	43.75	53.85
語彙	6	5	8	54.55	42.86	48.00
文法	125	116	35	51.87	78.12	62.34
表現	0	0	16	-	0	0
その他	0	0	7	-	0	0
計	152	130	93	53.90	62.04	57.69

⑧課題：誤り区間推定の今後

- ① 誤り文の単語分割
 - ▶ 「きれく」は辞書にない
 - ▶ 修正文に依存することがある
 - ▶ 「かわい|そ|だ」「かわいい|そう|だ」
 - ▶ 「かわいそ|だ」「かわいそう|だ」
- ② トップダウンの区間推定へ
 - ▶ ボトムアップ型 (ERRANT)
 - ▶ 単語単位の編集をマージ
 - ▶ 誤り区間の一貫性を保ちにくい
 - ▶ トップダウン型
 - ▶ 区間の候補を先に得て，選択

⑨課題：誤り種類推定の今後

- ① タグの設計
 - ▶ 下位の分類
 - ▶ 文法誤り
 - ▶ 品詞 (ERRANT 同様)
 - ▶ 表現誤り
 - ▶ 言い換え
 - ▶ モダリティ (話者の気持ち) の選択
 - ▶ 語用論的な誤り
 - ▶ 論理的な誤り
- ② 格関係とモダリティ
 - ▶ 命題とモダリティの対比に着目 (寺村 82)
 - ▶ 「本を読む可能性がある」
 - ▶ 「本を読むかもしれない」
 - ▶ 「本を読」までが「命題」で「格関係」
 - ▶ 後ろが「モダリティ」
 - ▶ 格関係
 - ▶ 語彙誤り・文法誤り
 - ▶ モダリティ
 - ▶ 文法誤り・表現誤り
 - ▶ 英語にはない発想が必要

補足ポスター

誤り種類の仕様はどう決めるか？

- ▶ James 98 の誤り分類法
 - ① 逸脱様態に基づく分類 (modification)
 - ▶ 脱落, 付加, 誤形成, 語順
 - ② 言語学的分類 (linguistic)
 - ▶ 表記, 文法, 語彙, 文章, 談話
 - ▶ 品詞・文法カテゴリ・同義語対義語・コロケーション等は, その下位分類になる
- ▶ Granger 03 の誤り種類の設計方針
 - ① 有益性・情報提供性 (informativity)
 - ▶ 学習者のために詳細に情報を提供できる
 - ② 運用性・管理容易性 (manageability)
 - ▶ アノテーターが運用できるよう詳細すぎない
 - ③ 再利性・再利用可能性 (reusability)
 - ▶ 異なるドメインでも一般的に利用できる
 - ④ 柔軟性 (flexibility)
 - ▶ 誤り種類の変更 (追加・削除) が容易
 - ⑤ 一貫性 (consistency)
 - ▶ アノテーター間での差異がなるべく小さい

参考文献

- C. Bryant, M. Felice, T. Briscoe (2017). “Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction.”
- C. James (1998). “Errors in Language Learning and Use: Exploring Error Analysis.” London and New York: Longman.
- S. Granger (2003). “Error-tagged Learner Corpora and CALL:A Promising Synergy.” CALICO Journal 20/3, 465–480.
- 吉川武時 (1997). 外国人の日本語誤用分析 誤用分析 I. 明治書院企画編集部編『日本語誤用分析』明治書院.
- A. Koyama, T. Kiyuna, K. Kobayashi, M. Arai, M. Komachi (2020). “Construction of an Evaluation Corpus for Grammatical Error Correction for Learners of Japanese as a Second Language.” LREC 2020.
- 宮田学 編 (2002). ここまで通じる日本人英語 新しいライティングのすすめ. 大修館書店.
- A. Lüdeling, M. Walter, E. Kroymann, P. Adolphs (2005). “Multi-level error annotation in learner corpora.”

誤り種類はどう決めたか？

- ▶ 吉川 97 の 4 分類 (言語学的分類) + その他
 - ▶ 表記・語彙・文法・表現・その他
- ▶ 運用性・再利性・柔軟性・一貫性を満たす

今後の課題として考えていること

- ① 誤り文の単語分割の改善
 - ▶ 誤り文の単語分割が失敗すると, 誤り区間・誤り種類の推定も失敗する
 - ▶ 「きれく」は辞書にない
 - ▶ 修正文の分割を参照する誤り文の分割は？
 - ▶ 誤り文の分割は, 訂正文の影響を受ける
 - ▶ 「かわい | そ | だ」→「かわいい | そう | だ」
 - ▶ 「かわいそ | だ」→「かわいそう | だ」
- ② ERRANTの枠組みでの誤り区間の推定の限界は？
 - ▶ 単語単位の編集をマージするボトムアップの枠組みは, 誤り区間に一貫性を持たせにくい
 - ▶ 基準に適した誤り区間の候補を先に得て, 選択するトップダウンの枠組みに可能性？
- ③ 評価手法の評価
 - ▶ より一貫性のあるアノテーション基準の作成
- ④ 有益性の高い誤り種類の設計
 - ▶ 品詞・文法カテゴリ等 → 誤用分析研究を参考
- ⑤ 「表現の誤り」や「流暢性」などの都合の良い概念に依存していないか？
 - ▶ 実質的に「その他」と差がなく, 有益性が低い
 - ▶ 「論理的誤り」「語用論的誤り」(宮田 + 02)
 - ▶ アノテーションスキームの見直し
 - ▶ Multi-layer standoff annotation (Lüdeling+ 05)
 - ▶ 日本語に対して, 高い表現力のアノテーションに利点があるか？(語順・談話などは?)
- ⑥ 係り受け構造の利用
 - ▶ 構造の違いは, 表現の違い
 - ▶ 表現の誤りは, 構造が手がかりになる