文法誤り訂正における訂正難易度の判別可能性

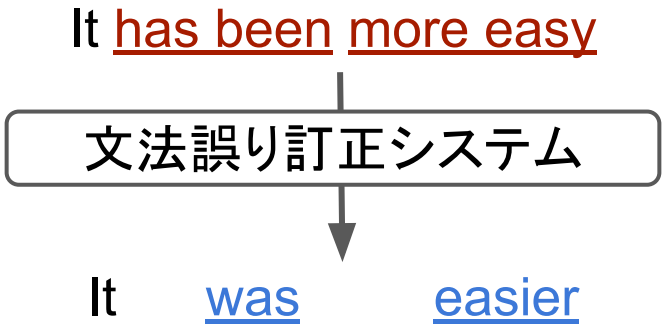
五藤巧[1)]，永田亮[2,3)]，三田雅人[3),a)]

1) 奈良先端科学技術大学院大学
2) 甲南大学
3) 理化学研究所

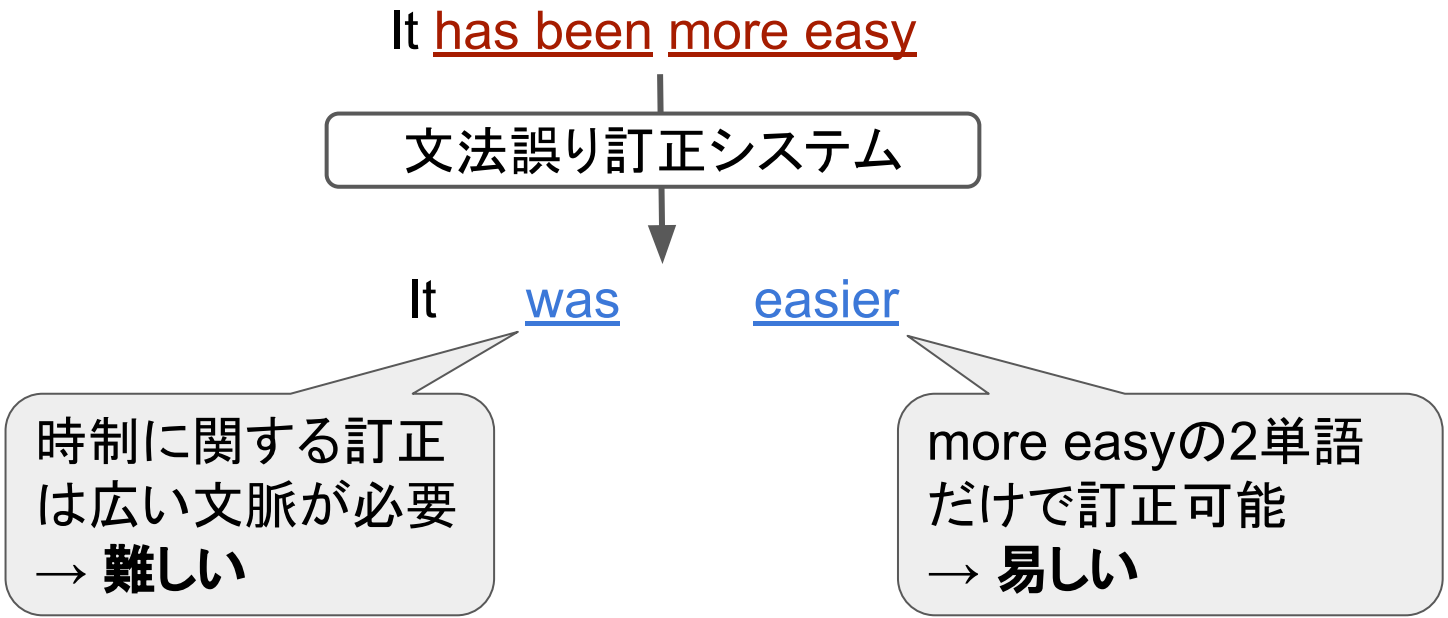
a)（現所属は株式会社サイバーエージェント）

文法誤り訂正の評価を考える

- 以下のような誤り訂正結果をどう評価するか？

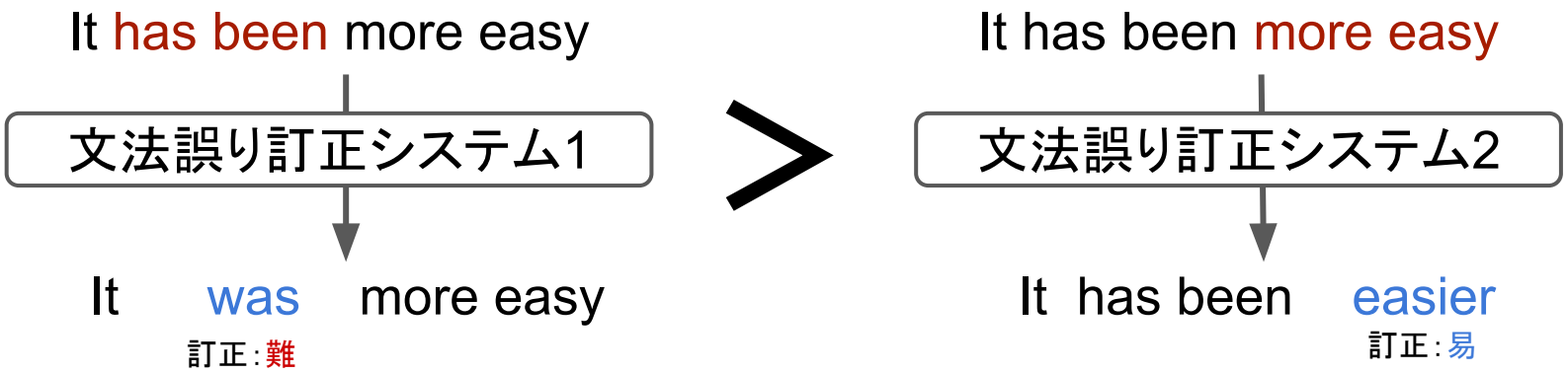


文法誤りには訂正難易度がある



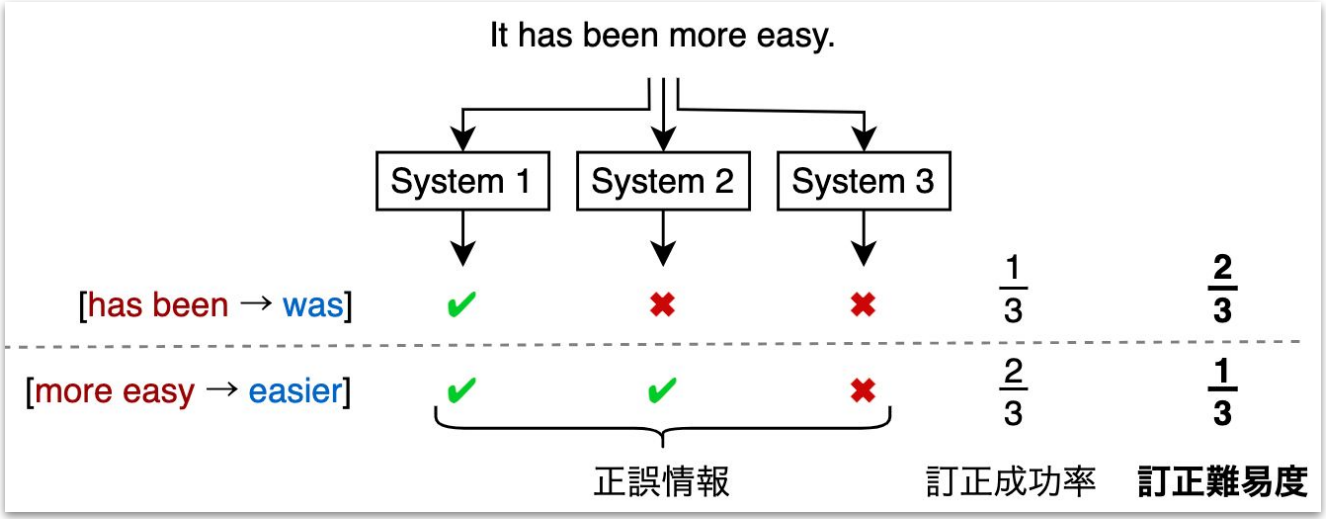
訂正難易度を考慮した評価[Gotou+ 20]

- 訂正が難しい誤りを訂正したシステムを高く評価
  → 研究コミュニティが**訂正が難しい誤りに挑戦する動機付け**に



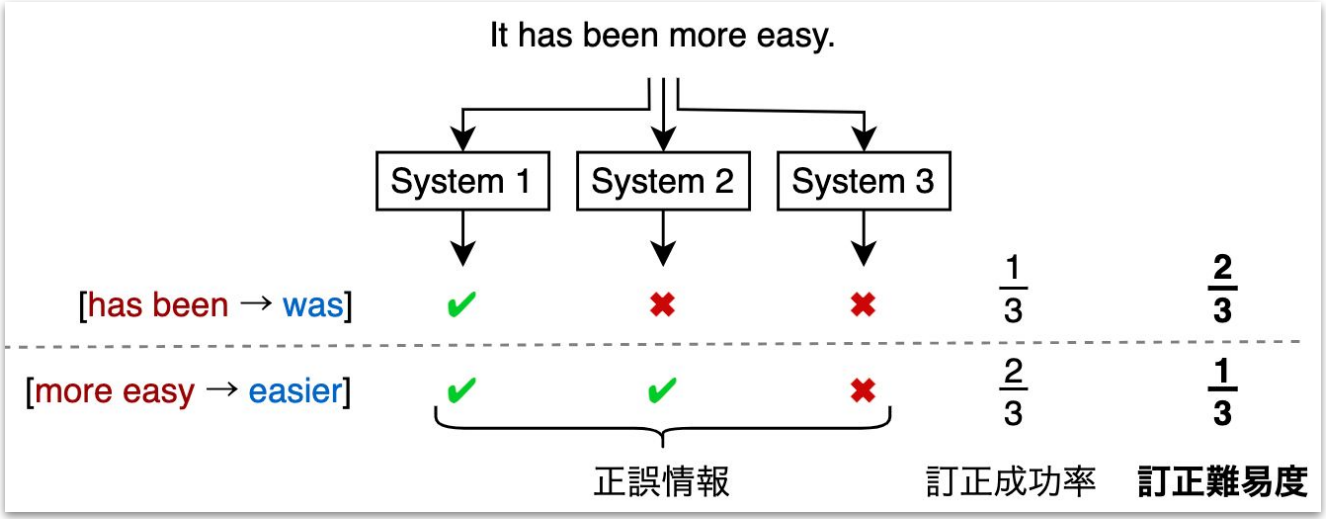
訂正難易度の自動計算[Gotou+ 20]

- **複数のシステム出力**から得られる**訂正成功率**から計算
  - 訂正難易度 = 1 - 訂正成功率



自動計算された訂正難易度に残る疑問点

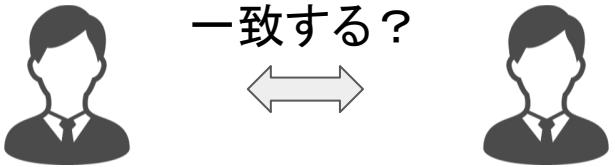
- **複数のシステム出力**から得られる**訂正成功率**から計算
  - 訂正難易度 = 1 - 訂正成功率

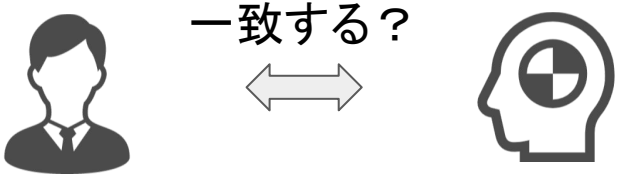


自動計算した訂正難易度は，人が判別する難易度を反映する？

本研究のリサーチクエスチョン

- そもそも，訂正難易度の判別が**人同士**で一致するのか？
- **人と機械**（=自動計算）では一致するのか？





人が判別する訂正難易度をどう調査する？

- **誤りのペアを提示**し，より訂正が**難しいほうを選択**
- 相対的な尺度で訂正難易度を判別するため，判別が容易

①かな！

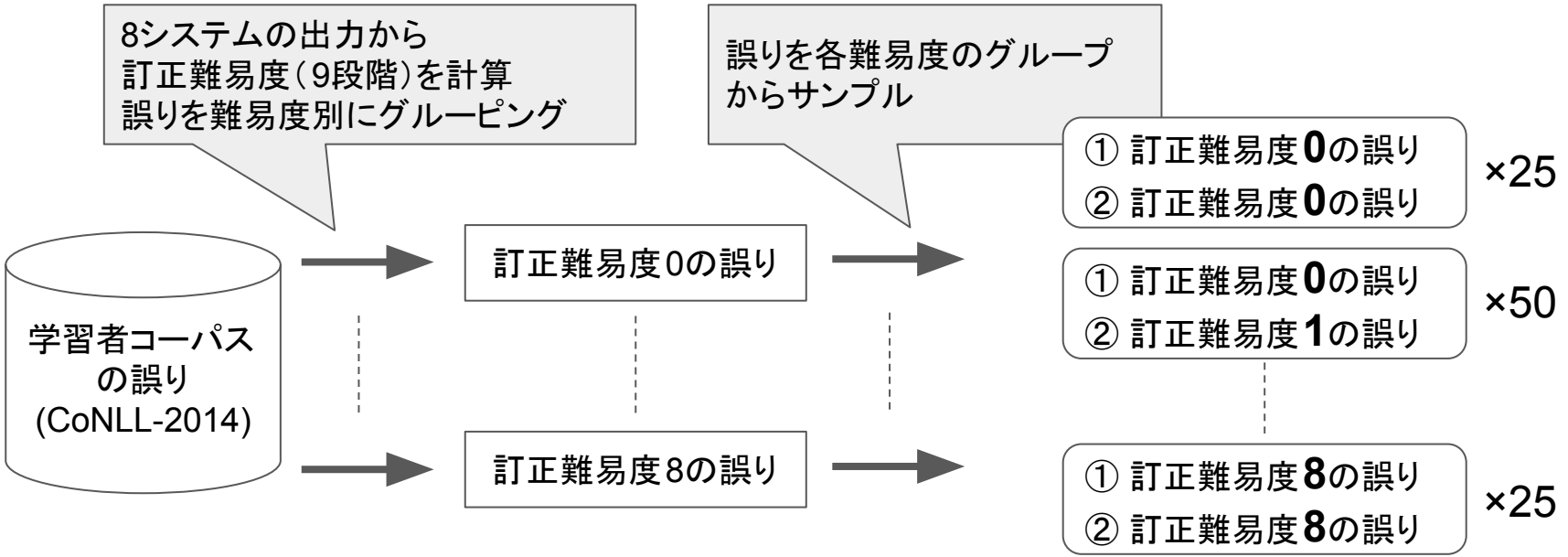
① It [has been → was] more easy.
② It has been [more easy → easier]

どっちの誤り訂正が難しいと思う？

判別者に提示する誤りペアの要件

- **多様**な訂正難易度のペアが**均等**に含まれるべき
  - 難易度の組み合わせによって一致に違いが出るかを分析したい
- ランダムサンプルでペアを作成するのは適さない

機械の訂正難易度を用いた誤りペア生成



- 2025件の誤りペアを作成
  - 異なる難易度のペアを50件ずつ，同じ難易度のペアを25件ずつ作成

難易度判別実験

- **判別者**
  - 第二，第三著者
- **手順**
  - 判別結果を4値でアノテーション
  - 独立に実施

誤り 1	判別結果	誤り 2
It is *more easy → easier to...	<	It is difficult for *the → ϕ ...
*A → The doctor said ...	>	The number *coresponds → corresponds ...
It can be *improve → improved ...	=	It can be *explain → explained ...
*Some how → Somehow I must find ...	? (判別不能)	*May be → Maybe I go to ...

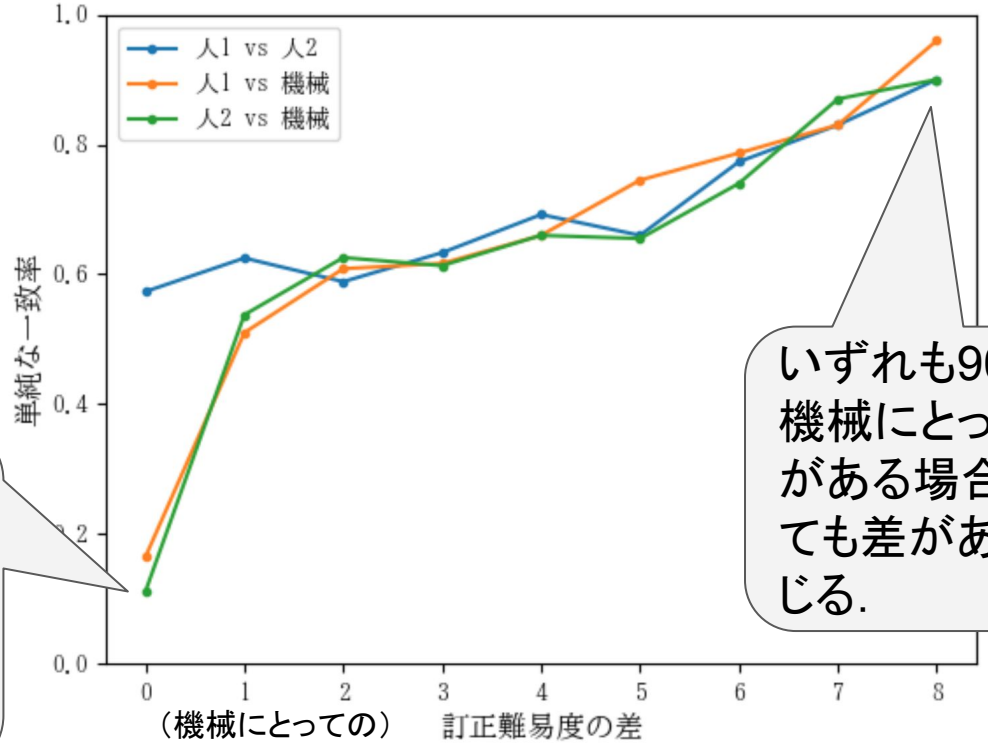
結果：全体的な一致率

- **人同士**はある程度一致
  - Majorityベースライン（46.3%）よりも優位に高い
  - 完全な一致からはほど遠いため，人にとっても判別は難しい
- **人と機械**も**人同士**に遜色ない一致
  - 機械の難易度は人の難易度に遜色ない

評価のペア	一致率（%）	Cohen's-κ
人$_1$ vs 人$_2$	66.39	0.42
人$_1$ vs 機械	64.72	0.37
人$_2$ vs 機械	64.28	0.33

（同一難易度ペアを除いた場合）

難易度の“差”が広がるほど一致率が向上する



人は機械よりも細かい粒度で難易度を判別.
機械にとっては同じ難易度でも，人にとっては差があるように感じるため一致率が低下

いずれも90%以上.
機械にとって大きな差がある場合，人にとっても差があるように感じる.

誤りタイプ別に見た一致

- 全体的に一致の傾向は類似
- 機械判別が難しい誤りは人にも難しい
  - **ADJ**: 難しいと判断されて一致
- 機械判別が易しい誤りは人にも易しい
  - **VERB:INFL**: 易しいと判断されて一致

	単純な一致率		
誤りタイプ	人$_1$ vs 機械	人$_2$ vs 機械	人$_1$ vs 人$_2$
ADJ	0.92	0.85	0.92
VERB:INFL	0.90	1.00	0.90
WO	0.88	0.75	0.69
CONJ	0.87	0.73	0.73
PUNCT	0.75	0.80	0.81
NOUN	0.73	0.72	0.58
ADV	0.72	0.72	0.72
VERB	0.72	0.68	0.70
PART	0.70	0.63	0.59
ADJ:FORM	0.69	0.46	0.54
VERB:FORM	0.69	0.65	0.68
PRON	0.67	0.72	0.72

不一致の分析：訂正アルゴリズム依存の側面

- **ORTH**（空白に関する結合・分割誤り, e.g. *some how → somehow*）
  - 人同士で不一致
  - **人1**：システムは空白により分割された単語単位・サブワード単位で扱う
    → 訂正は難しい
  - **人2**：専用の規則と辞書引きによって訂正可能
    → 訂正は易しい

	単純な一致率		
誤りタイプ	人1 vs 機械	人2 vs 機械	人1 vs 人2
ORTH	**0.45**	**0.70**	**0.30**

不一致の分析：訓練データ依存の側面

- **SPELL**（綴り誤り）
  - 人と機械で不一致
  - **機械**：サブワード単位で扱ったとしても，訓練データになければ訂正できない可能性が高い
  - **人**：訓練データを考慮した難易度判別は困難

	単純な一致率		
誤りタイプ	人1 vs 機械	人2 vs 機械	人1 vs 人2
SPELL	**0.54**	**0.61**	**0.77**

人および機械の訂正難易度をどう使うべき？

- **人が判別する訂正難易度**
  - 少量なデータに対する詳細な分析に利用可能
  - ✅ 機械よりも詳細な粒度で難易度を判別可能
  - ❌ 時間と労力といったコストが高い
- **機械が判別する訂正難易度**
  - Shared Taskのような大規模な環境での難易度分析に利用可能
  - ✅ 大量の誤りに短時間で難易度を付与可能
  - ✅ 人でも気がつかない“訂正の難しさ”の発見につながる
  - ❌ 人よりも難易度の粒度は粗い

まとめ

- **背景**：訂正難易度を考慮した評価は重要であるが，自動計算された難易度は人が判別する難易度を反映しているか明らかでない
- **リサーチクエスチョン**
  - そもそも，訂正難易度の判別が人同士で一致するのか？
    → ある程度一致（66%）
  - 人と機械では一致するのか？
    → 人同士に遜色ない一致（64%）

参考文献

- Takumi Gotou, Ryo Nagata, Masato Mita and Kazuaki Hanawa, **“Taking the Correction Difficulty into Account in Grammatical Error Correction Evaluation”**, In Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)

付録:（p.9の）8システムの詳細

- 6種のNNベースモデルと2種のSMTベースモデル
  - [Kiyono+ 2019]: Transformer-based
  - [Junczys-Dowmunt and Grundkiewicz+ 2016]: SMT-based
  - [Ge+ 2018] : CNN-baed
  - [Junczys-Dowmunt+ 2018]: Transformer-baed
  - [Mita+ 2019]の4つのベースラインモデル
    - LSTM
    - SMT
    - CNN
    - Transformer

付録:（p.19の）8システムの詳細，参考文献

- Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. **An empirical study of incorporating pseudo data into grammatical error correction**. In Proceedings of EMNLP-IJCNLP, pages 1236– 1242.
- Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. **Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction**. In Proceedings of EMNLP, pages 1546–1556.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. **Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task**. In Proceedings of NAACL, pages 595–606.
- Tao Ge, Furu Wei, and Ming Zhou. 2018. **Reaching Human-level Performance in Automatic Grammatical Error Correction: An Empirical Study**. arXiv.
- Masato Mita, Tomoya Mizumoto, Masahiro Kaneko, Ryo Nagata, and Kentaro Inui. 2019. **Cross-corpora evaluation and analysis of grammatical error correction models — is single-corpus evaluation enough?** In Proceedings of NAACL: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1309–1314, June.

付録:(p. 11の)Majorityベースラインとは

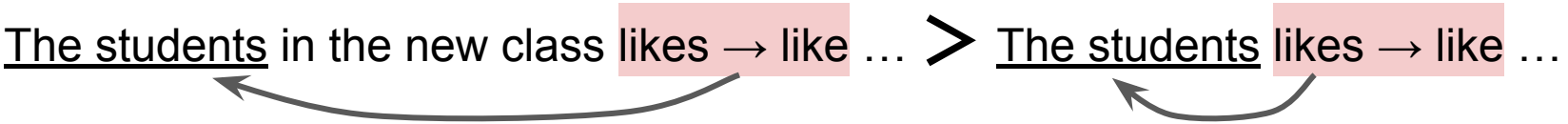
- 第一判別者が全ての評価結果を < とした場合の第二判別者との一致率
  - 938 / 2025 = 0.463 (46.3%)

表 4　判別結果の分布.

評価者	比較結果			
	<	=	>	?
人 1	922	256	846	1
人 2	938	167	920	0

付録:難易度判別実験における基準

- **基準1:** 訂正に広い文脈が必要であるほど難しい



- **基準2:** 語彙に関わる訂正は難しい
- **基準3:** 複合的な誤り訂正は難しい

A students likes → like … > Students likes → like …

一致率の計算は容易

- 判別結果を比較するだけ