ソフトな語彙制約による テキスト難易度制御の検討

西原大貴[1]　梶原智之[2]　荒瀬由紀[1]

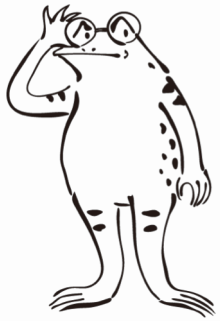
[1]大阪大学情報科学研究科　[2]大阪大学データビリティフロンティア機構

テキストの難易度制御

難しいテキスト

テキスト平易化システム

中学生向けのテキスト



小学生向けのテキスト

テキストの難易度制御の必要性

- 言語学習の教育現場では、
  学習者に適した難易度のテキストが必要

- 教師が各学習者向けに人手で平易化している
- 負担軽減のために自動化が求められている

テキスト平易化

- 難解な入力文を含意する平易な文を生成

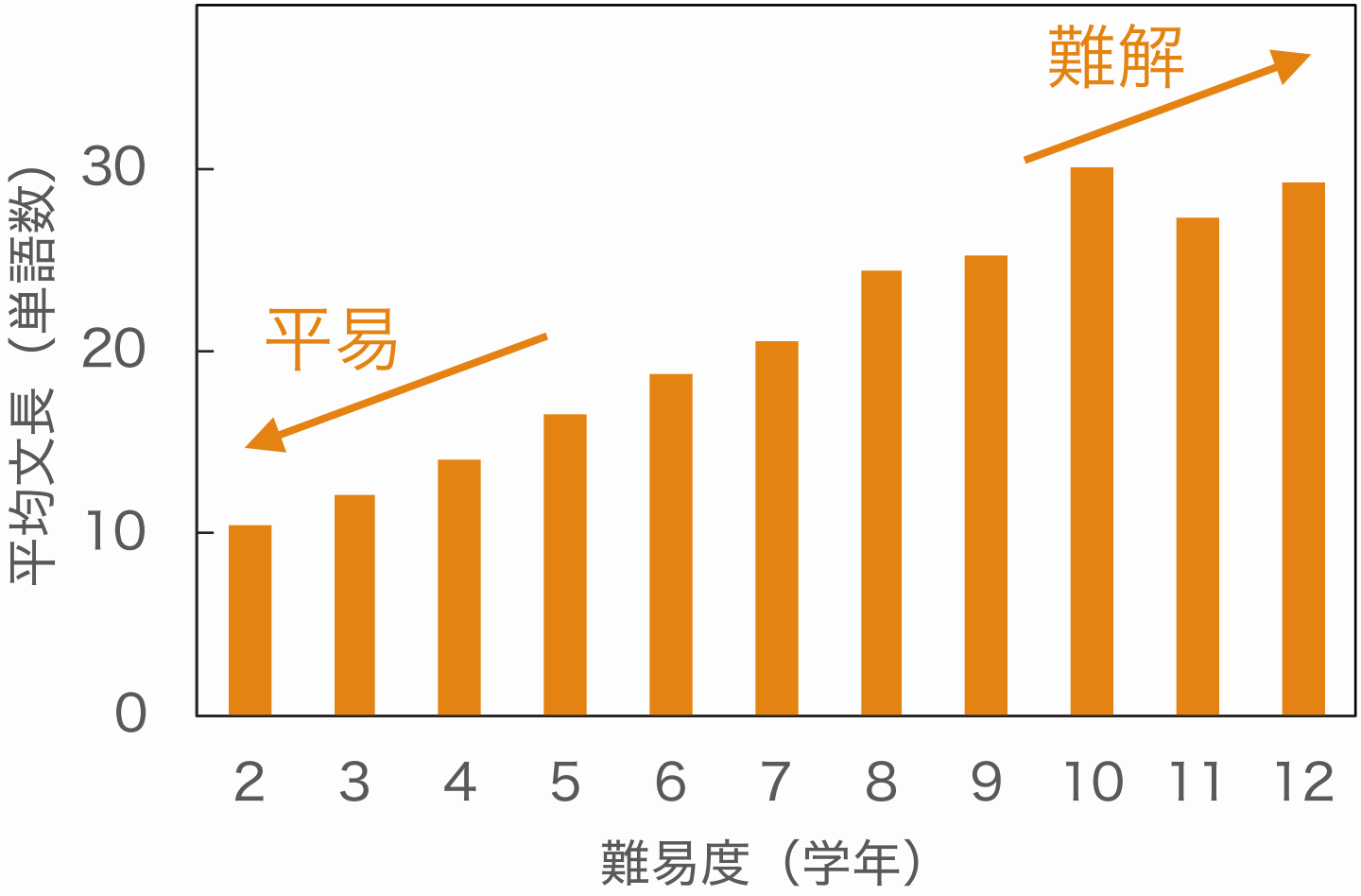
学年	例
入力 12	According to the Pentagon , 152 female troops have been killed while serving in Iraq and Afghanistan .
7	The Pentagon says 152 female troops have been killed while serving in Iraq and Afghanistan .
5	The military says 152 female have died .

テキスト平易化

- 難解な入力文を含意する平易な文を生成
- テキスト平易化の代表的な操作：置換と省略

学年	例
入力 12	According to the Pentagon , 152 female troops have been killed while serving in Iraq and Afghanistan .
7	The Pentagon says 152 female troops have been killed while serving in Iraq and Afghanistan .
5	The military says 152 female have died .

平易文は短い



テキスト平易化の既存手法

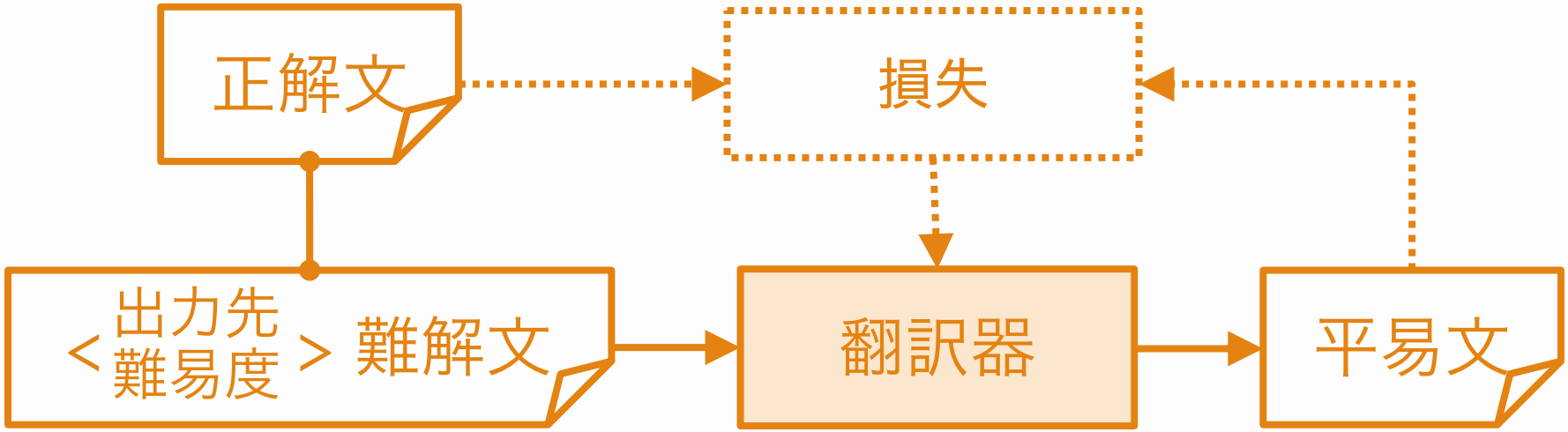
- テキスト平易化は同一言語内の翻訳問題

- 多くの研究[2-6]では２段階の平易化のみ
- あまり書き換えない保守的なモデルになる[3]

難解文	According to the Pentagon , 152 female troops have been killed while serving in Iraq and Afghanistan .
平易文	The Pentagon says 152 female troops have been killed while serving in Iraq and Afghanistan .

[2] Nisioi et al.:Exploring Neural Text Simpli cation Models. In Proc. of ACL, Vol. 2, pp. 85-91 (2017).
[3] Zhang and Lapata: Sentence Simpli cation with Deep Reinforcement Learning. In Proc. of EMNLP, pp. 584-594 (2017).
[4] Vu et al.: Sentence Simpli cation with Memory-Augmented Neural Networks. In Proc. of NAACL, Vol. 2, pp. 79-85 (2018).
[5] Guo et al.: Dynamic Multi-Level Multi-Task Learning for Sentence Simplification. In Proc. of COLING, pp. 462-476 (2018).
[6] Zhao et al.: Integrating Transformer and Paraphrase Rules. In Proc. of EMNLP, pp. 3164-3173 (2018)

既存手法：文の難易度を考慮[7]



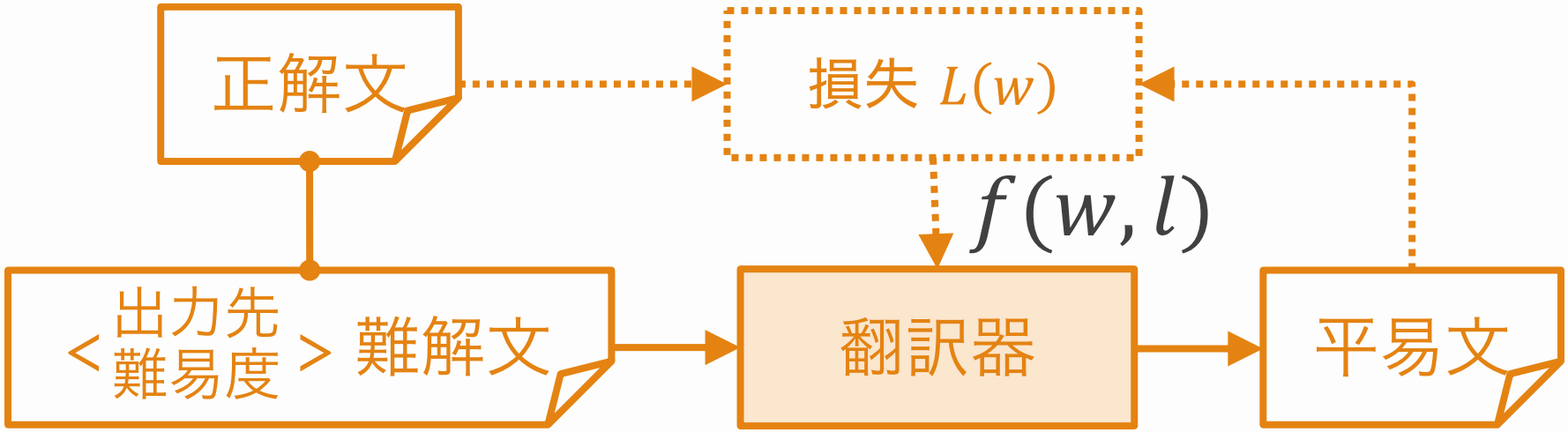
<5> According to the Pentagon , female troops have been killed while serving in Iraq .

The Pentagon says female troops have been killed .

- 入力文頭に平易文の難易度を付与する
- 文の難易度を考慮する：省略は得意
- 出力文に難解な単語がしばしば現れる

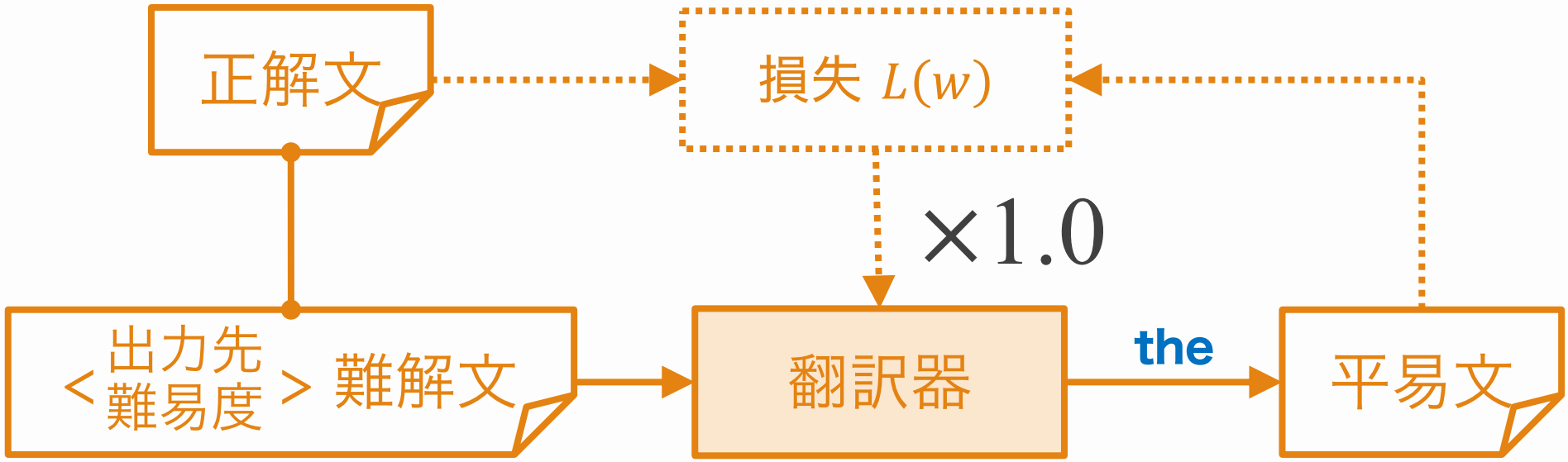
[7] Scarton and Specia: Learning Simplifications for Specific Target Audiences, In Proc. of ACL, pp. 712–718 (2018).

提案手法：単語の難易度も考慮



- 仮定：易しい単語は易しい文中で出現しやすい
- 各単語のクロスエントロピー損失 $L(w)$ を
  文の難易度 l に対する単語 w の出現しやすさ
  $f(w,l)$で重み付け

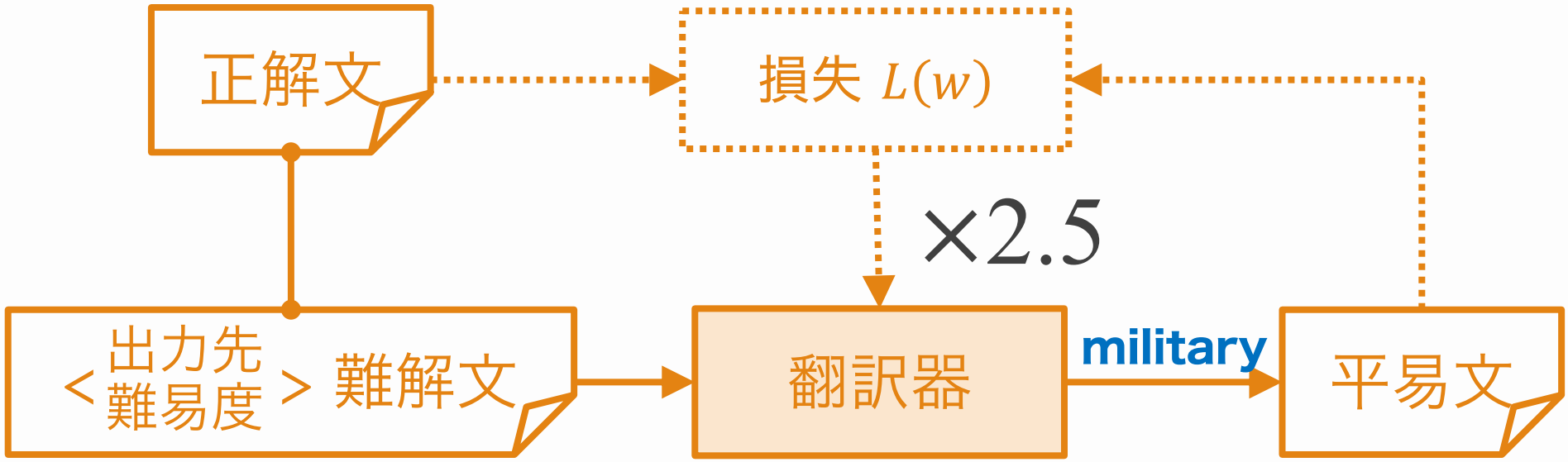
“the”はどの難易度でもよく出る



入力：
<5> According to **the** Pentagon , 152 female troops have been killed .

出力：
The military says 152 female have died .

“military”は**<5>**でよく出る



入力：
<5> According to the **Pentagon** , 152 female troops have been killed .

出力：
The **military** says 152 female have died .

重み $f(w,l)$：TFIDF

- 特定の難易度の特徴的な単語に重み付け

$$\mathrm{TFIDF}(w,l) = P(w|l)\log\frac{D}{DF(w)}$$

- $P(w|l)$は特定の難易度 l に対する単語 w の出現確率
- $D = 11$ は難易度の総数
- $DF(w)$ は単語 w が出現する難易度の種類数

重み $f(w,l)$：PPMI

- 文の難易度と単語の共起の強さPMI

$$\mathrm{PMI}(w,l) = \log \frac{P(w,l)}{P(w)P(l)} = \log \frac{P(w|l)}{P(w)}$$

- 重みは正数でないといけない

$$\mathrm{PPMI}(w,l) = \max(\mathrm{PMI}(w,l), 0)$$

$$f(w,l) = \mathrm{Func}(w,l) + 1, \qquad \mathrm{Func} \in \{\mathrm{TFIDF}, \mathrm{PPMI}\}$$

データセット：Newsela

- 難易度 2〜12（米国学校制度の学年）が
  専門家によって文書ごとに付与されている
- Xu et al. [8] のアライメント
- Zhang and Lapata [3] の分割

訓練用	1,070 文書	94,208 文対
検証用	30 文書	1,077 文対
評価用	30 文書	1,129 文対

[8] Xu et al., C.: Problems in Current Text Simplification Research: New Data Can Help, TACL, Vol. 3, pp. 283–297 (2015).
[3] Zhang and Lapata: Sentence Simplification with Deep Reinforcement Learning, In Proc. of EMNLP, pp. 584–594 (2017).

学習設定

- Marianを用いて以下の設定で学習
- 初期値を無作為に変更して 3 回の平均をとる

符号化器 復号器	２層の Bi-LSTM
隠れ層	1024次元
埋め込み層	512次元・dropout率0.1
最適化	Adam
early-stopping	Perplexity, 8 epochs

比較手法

- **s2s**
  難易度制御を行わないモデル
- **s2s+grade**
  既存手法（文の難易度を考慮）[7] の再実装
- **s2s+grade+TFIDF**
  提案手法（TFIDFで単語の難易度も考慮）
- **s2s+grade+PPMI**
  提案手法（PPMIで単語の難易度も考慮）

[7] Scarton and Specia: Learning Simplifications for Specific Target Audiences, In Proc. of ACL, pp. 712–718 (2018).

実験項目

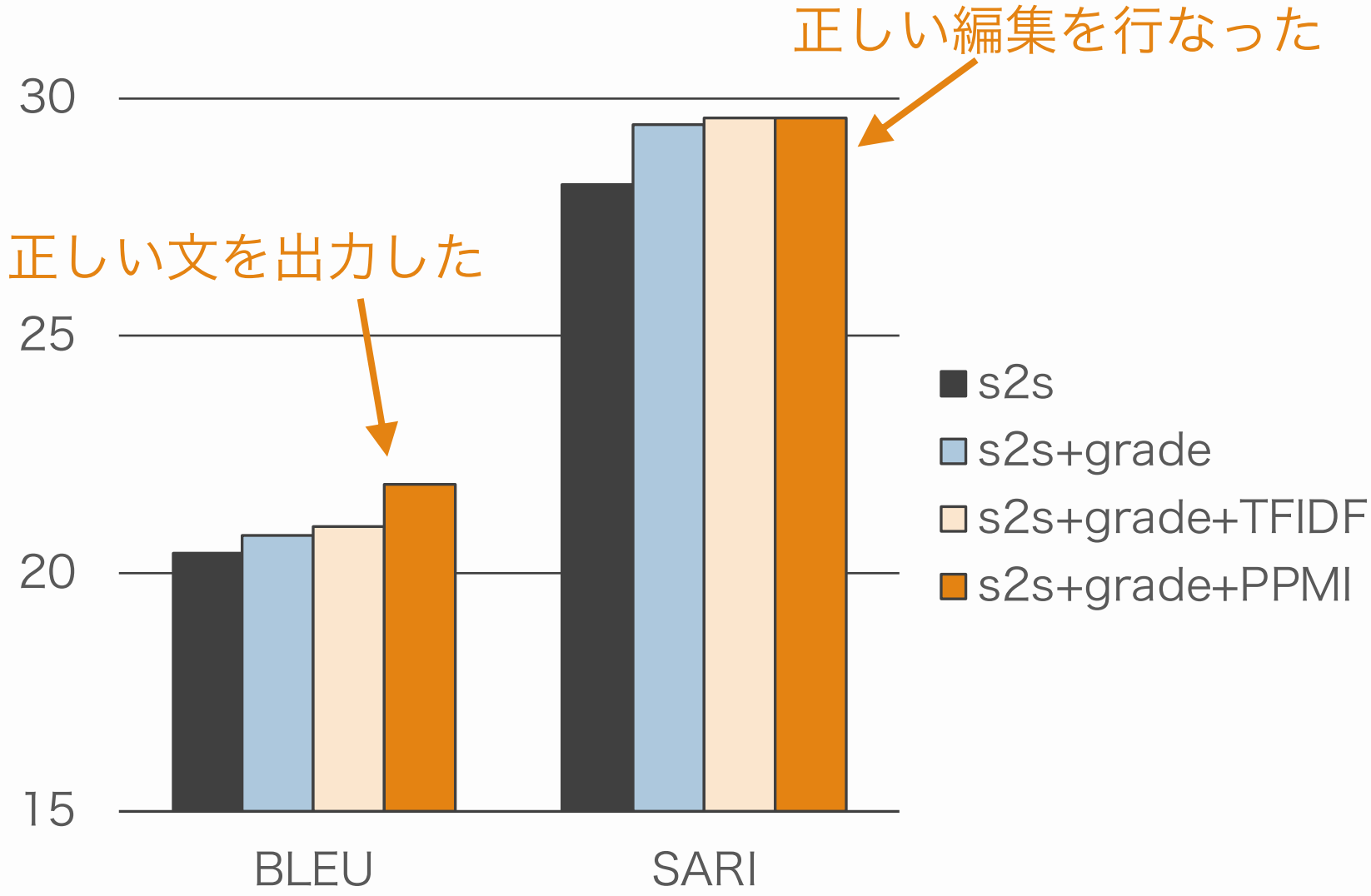
- **総合的な評価**
- 難易度ごとの分析
- エラー分析

総合的な評価

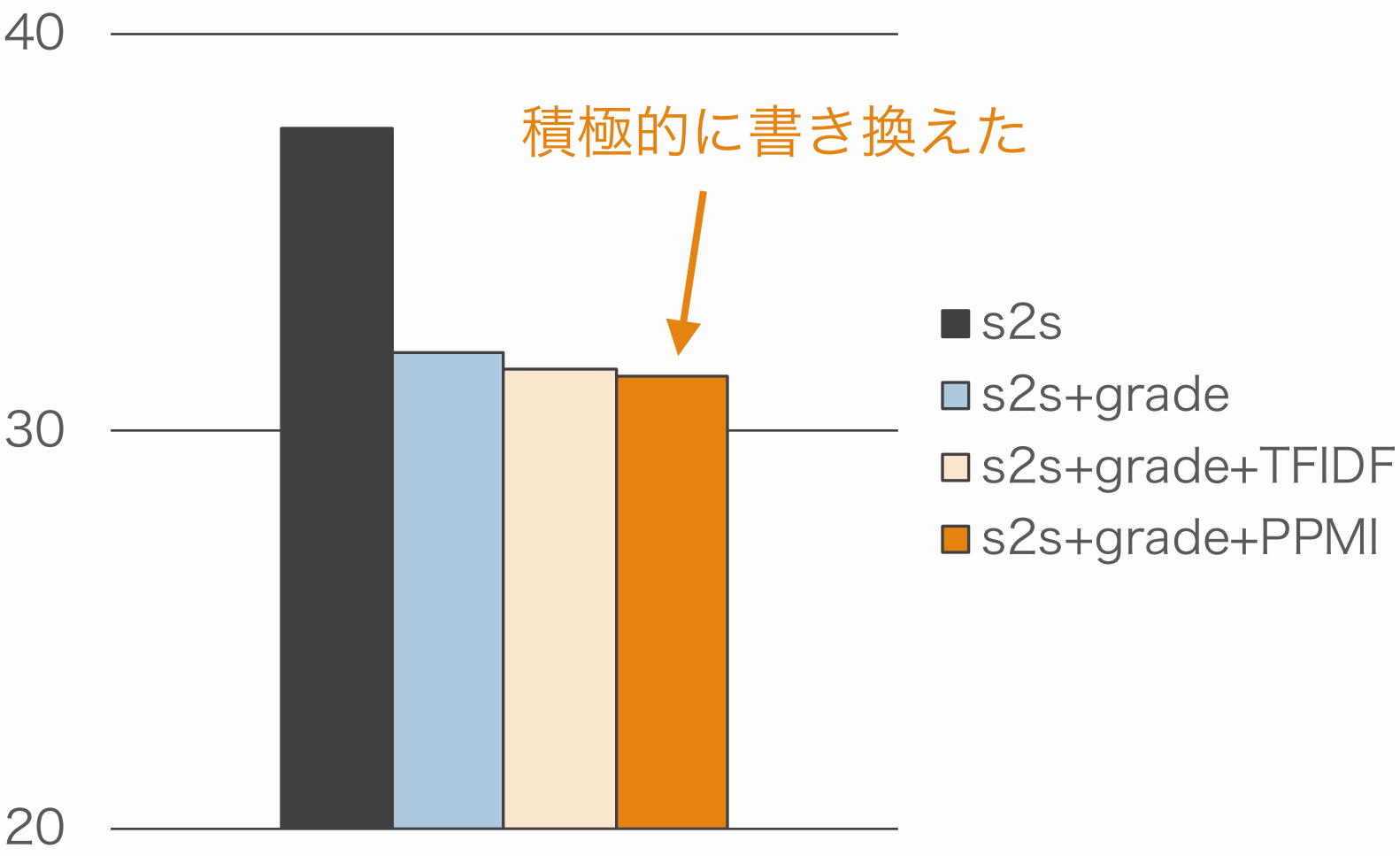
- BLEU↑：出力と正解の語句の一致率
- SARI↑：正しく追加・削除・保持したか

- $\mathrm{BLEU}_{\mathrm{ST}}$↓：入力と出力の語句の一致率
- 文長のMAE↓：文長の平均絶対誤差
- MPMI↑：出力単語の平均 PMI

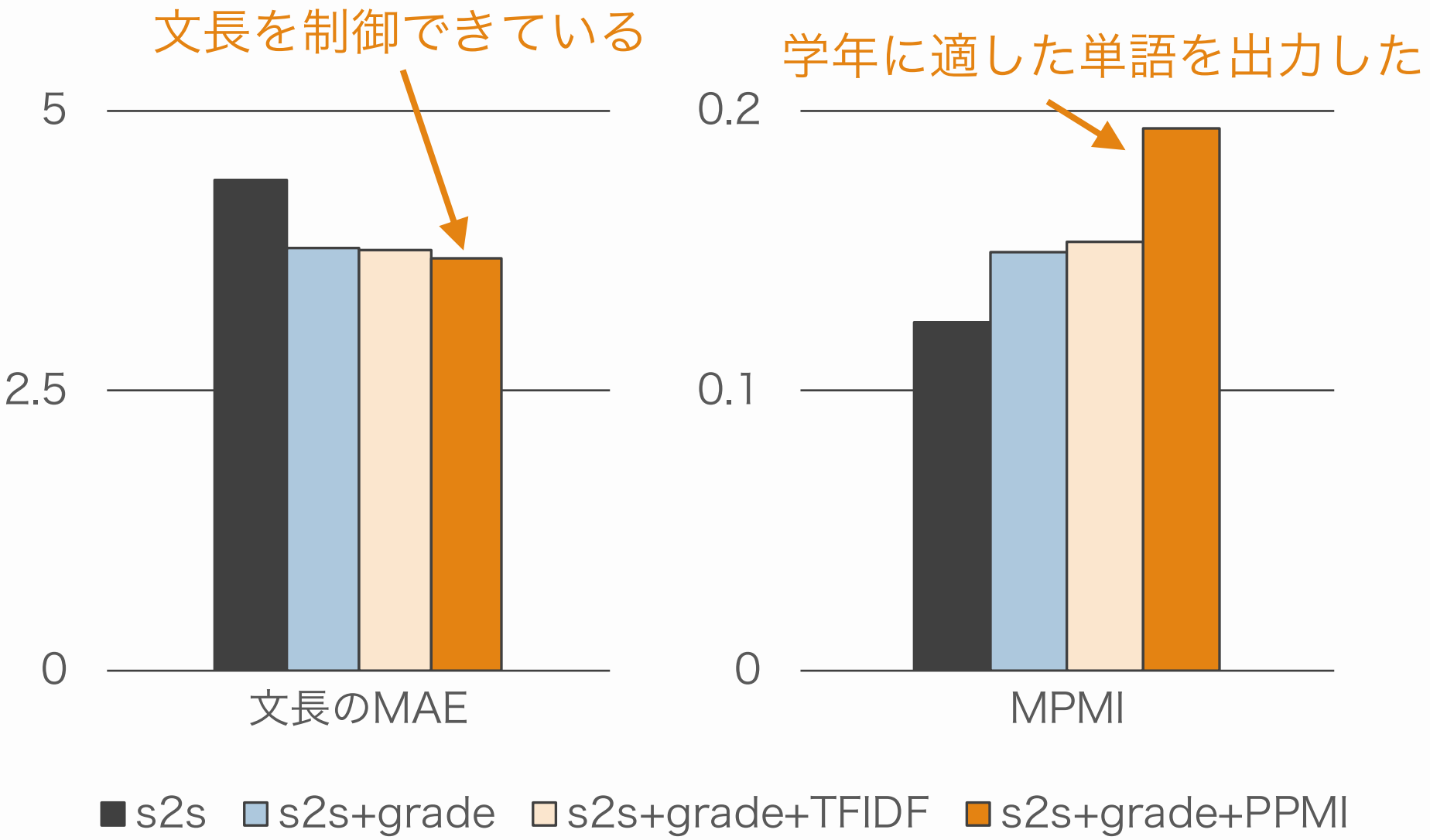
結果：BLEU↑・SARI↑


正しい編集を行なった
正しい文を出力した
30
25
20
15
BLEU
SARI
s2s
s2s+grade
s2s+grade+TFIDF
s2s+grade+PPMI

結果：$BLUE_{ST}$ ↓


40
30
20
積極的に書き換えた
s2s
s2s+grade
s2s+grade+TFIDF
s2s+grade+PPMI

結果：文長のMAE↓・MPMI↑



実験項目

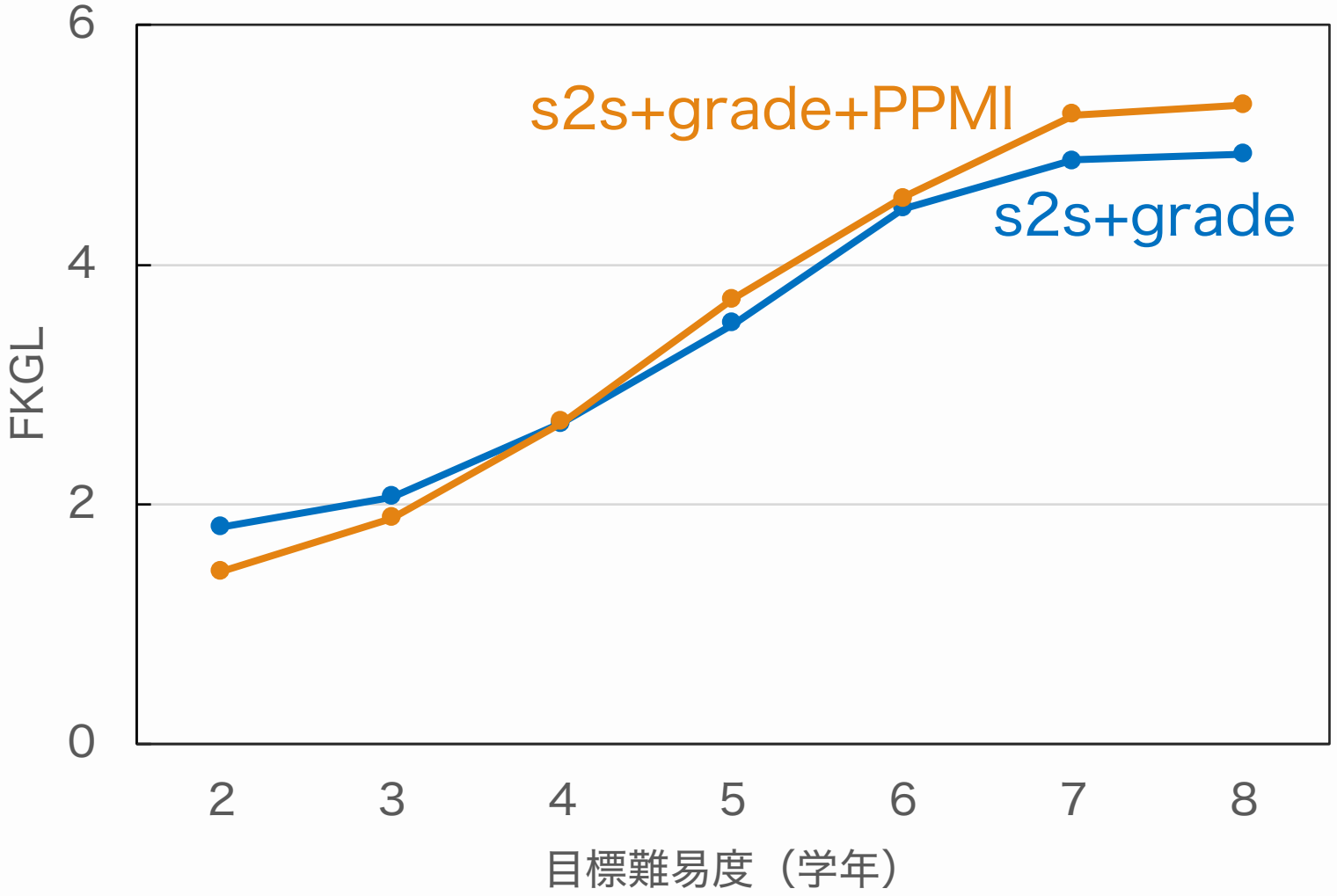
- 総合的な評価
- **難易度ごとの分析**
- エラー分析

難易度ごとの分析

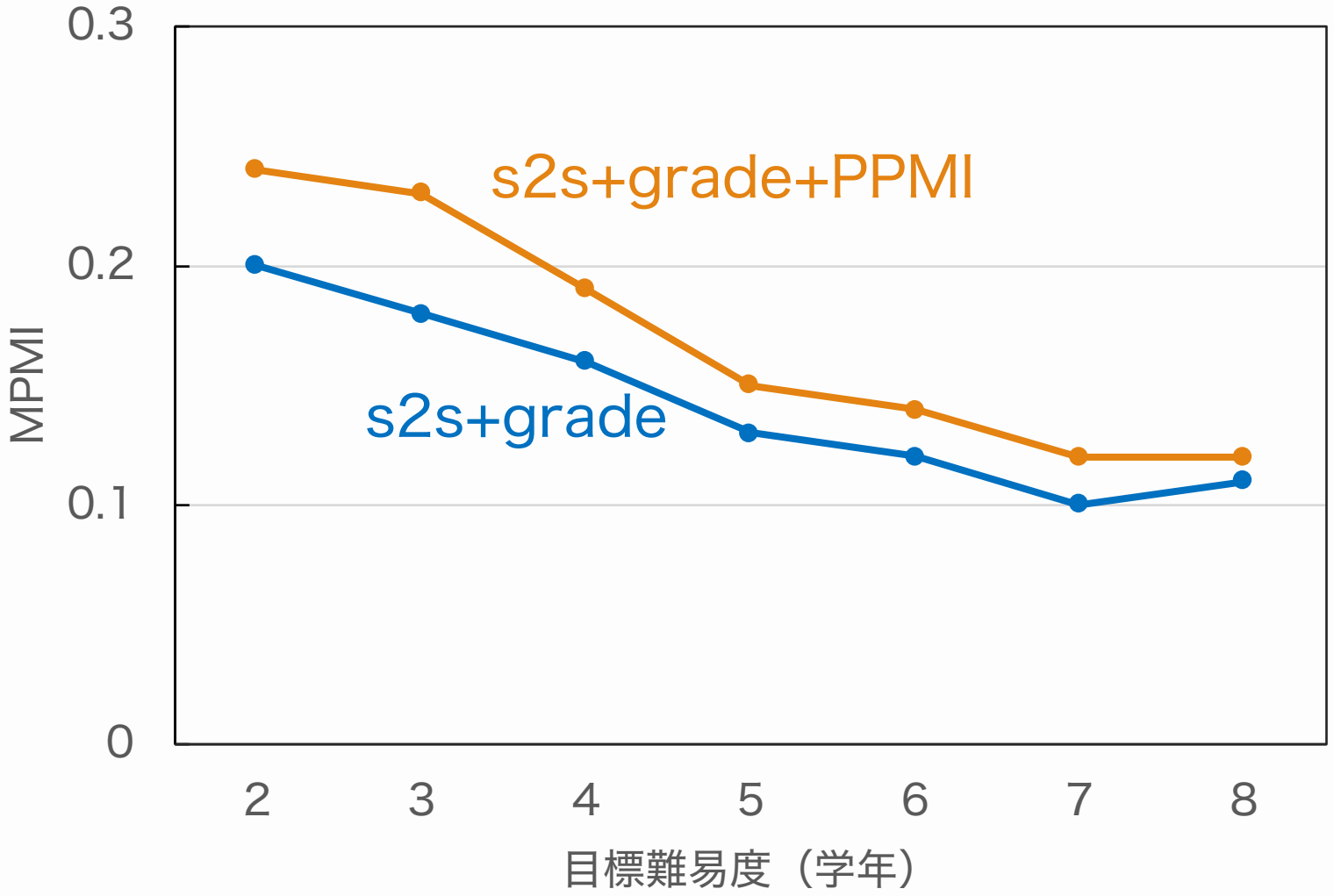
目標難易度ごとに以下の２つを評価

- FKGL
  - テキストの可読性指標
  - 易しいほどスコアが低い
- MPMI↑
  - 目標難易度に適した単語を出力できたか

メリハリがついた



目標難易度の単語が出力できた

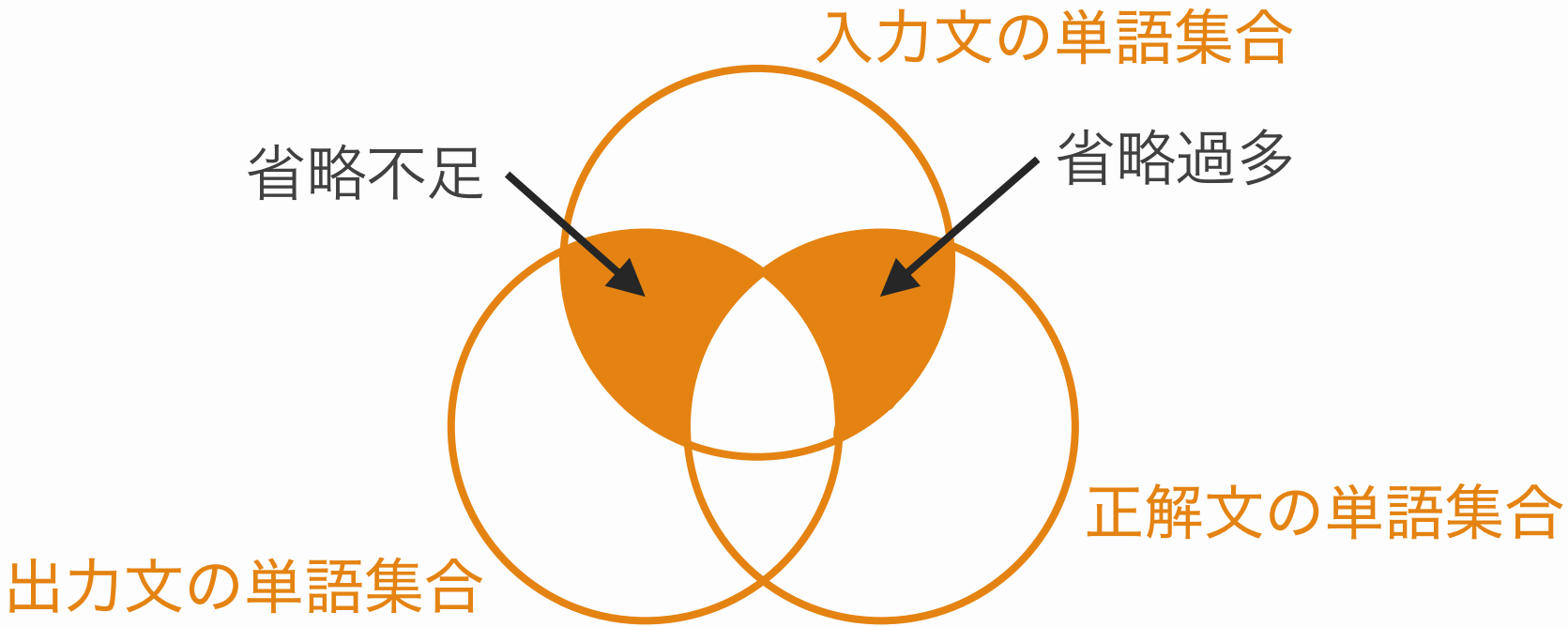


実験項目

- 総合的な評価
- 難易度ごとの分析
- **エラー分析**
  - 無作為抽出した25文について、
    省略と置換のエラーを数えた

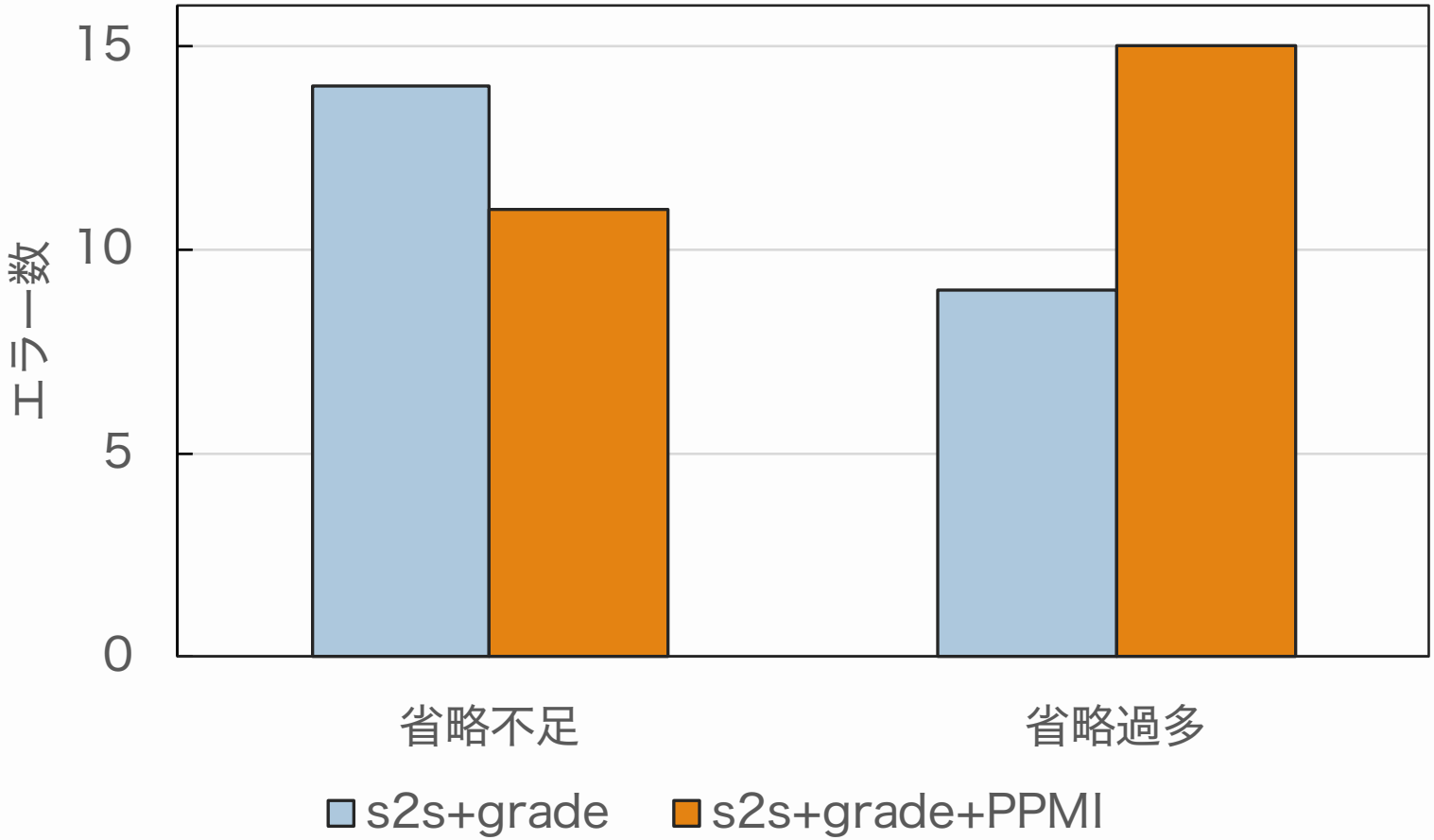
エラー分析：省略

- 省略不足：省略すべきだが省略しなかった箇所
- 省略過多：入力から残すべきだが省略した箇所



分析結果：省略

提案手法の方が積極的に省略している



エラー分析：置換

- 置換すべき箇所
- 置換成功：正しく置換した箇所
- 置換失敗：誤った表現に置換した箇所
- 置換不足：置換すべきだが置換しなかった箇所
- 置換過多：置換すべきでないが置換した箇所

入力文	She said the college application process can be especially stressful for immigrant students .
正解文	She said the college application process can be especially tough for immigrants .
出力文	She said the college application process can be very hard for immigrant students .

分析結果：置換

	既存手法	提案手法
$再現率 = \frac{置換成功}{置換すべき場所}$	4.0%	8.3%
$適合率 = \frac{置換成功}{置換成功＋置換失敗＋置換過多}$	4.7%	8.7%
$網羅率 = \frac{置換成功＋置換失敗}{置換すべき箇所}$	44.0%	56.3%

分析結果のまとめと今後の課題

- 置換の再現率も適合率も低い
- そもそも半分しか書き換えようとしていない
- コピーや省略に比べて、置換は難しい
- 翻訳は置換のみなのでリスクを取るしかないが
  テキスト平易化ではコピーや省略などの
  低リスクな操作に甘えがち
- 入力に無い単語の出力に対して報酬を与えたい
- もっと大きなコーパスが必要かもしれない

まとめ

- 学習者支援のため難易度を細かく制御
- 既存手法は、文の難易度を考慮
- 提案手法は、単語の難易度も考慮
  - 単語の難易度に基づいて損失を重み付け
  - BLEU と SARI が改善
  - 目標難易度に適した単語を出力
- 今後の課題：入力にない単語に報酬を与える