文法性・流暢性・意味保存性に基づく文法誤り訂正の参照なし評価

- URL
  - https://www.jstage.jst.go.jp/article/jnlp/25/5/25_555/_article/-char/ja/
- 背景
  - [1]は参照文を使わず訂正文の文法性に基づき訂正を評価する手法を提案した。しかし参照有り手法であるGLEUを上回る性能での評価は実現できなかった。そこで本論文では[1]の参照無し手法を拡張し、その評価性能を調べる。具体的には、[1]が用いた文法性の観点に加え、**流暢性**と**意味保存性**の3 観点を考慮する組み合わせ手法を提案する。流暢性はGEC システムの出力が英文として**どの程度自然であるか**という観点であり、意味保存性は**訂正前後で文意がどの程度保たれているか**という観点である。
- 自動評価尺度の評価方法
  - 自動評価尺度に求められる性質のうち最も重要なものは、人手評価との相関が高いことであるとされている
    - 例：機械翻訳の評価尺度のshared taskであるWMT 2017 Metrics Shared Task
      - 評価
        - 翻訳システム単位
          - 人手評価によるシステムに対する評価と自動評価尺度によるシステムに対する評価を比べることで評価する
          - ピアソンの相関係数
          - スピアマンの順位相関係数
        - 文単位
          - システムの翻訳ごとに人手で優劣が付けられており、自動評価尺度によってその優劣を識別できるかで評価する

        - ケンドールの順位相関係数
- 既存の評価尺度
    - 機械翻訳
        - BLEU - 自動評価尺度
- 参照あり手法

  参照有り手法では正確な評価のために、各入力文に対する参照文を１個だけでなく複数個用いることができる。参照文を複数用いる場合、各文の評価はM^2およびI-measureでは最大値が採用され、GLEU は平均値が採用される。

    - M^2
        - 文法誤り訂正の初期の研究では、訂正システムが行った編集操作がどの程度正解の編集と一致しているかをF 値で評価していた。しかし、長いフレーズの編集が必要な場合などに訂正システムを過小評価してしまうという問題があった。この問題を解決するためにM^2は”edit lattice”を用いることにより、システムが行った編集操作を正解と最大一致するように同定する
    - I-measure
        - 上述のM2 の問題点として、訂正を全く行わないシステムと誤った訂正をしたシステムに対するスコアがどちらも0 となる点が挙げられる。そこで、入力文が改善されれば正の値、悪化すれば負の値をとる尺度であるI-measure が提案された。I-measure は入力文、訂正文、参照文に対してトークンレベルでアライメントを行い、精度（accuracy）に基づきスコアを計算する
    - GLEU
        - 機械翻訳の標準的な評価尺度であるBLEUをGEC のために改善した評価尺度である。GLEUは訂正文(H)と参照文(R)で一致するn-gram数から、原文(S)に現れるが参照文に現れないn-gram数を減算することによって計算される。形式的には次式で表される。

$$\mathrm{GLEU+} = \mathrm{BP} \cdot \exp(\sum_{n=1}^{4} \frac{1}{n} \log(p_n\prime))$$

$$p_n\prime = \frac{N(H,R) - [N(H,S) - N(H,S,R)]}{N(H)}$$

ただし、N(A,B,C,...)は集合間でのn-gram重なり数を表し、BPはBLEUと同様のbrave penaltyを表す。brave penaltyは入力文に対して出力文が短い場合にn-gram適合率を減点する項である。これまでに提案された参照有り手法の中では最も人手評価との相関が高い

- 参照なし手法
  - 機械翻訳
    - 品質推定
      - 参照文を用いずに翻訳の品質を評価する
      - 翻訳システムの出力の良さを測るために、Human-targeted Translation Error Rate (HTER)と呼ばれる、人間の翻訳とシステムの翻訳の編集距離がどの程度近いかを計算する指標が用いられる。
- 提案手法
  - 文法性、流暢性、意味保存性の３つの観点を考慮した参照無し評価手法
  - 文法性
    - GECシステムの出力に標準英語上の文法誤りがあるかどうかという観点
    - 文法性のスコアS_G(h)は言語学的な素性に基づくロジスティック回帰により求める
      - 素性
        - スペルミス数
        - n-gram言語モデルスコア
        - out-of-vocabulary数
        - PCFG
        - リンク文法

        - 依存構造解析に基づく数の不一致素性
        - Language Toolによる誤り検出数
    - 流暢性
        - システムの出力がどの程度自然な英文であるかという観点
        - 訂正文hに対し、流暢性S_F(h)を以下のように求める。|h|は文長、P_mは言語モデルによる生成確率、P_nはユニグラム生成確率
    - 意味保存性
        - 意味がどの程度保存されているかを測る単純な方法は，原文の単語が訂正後の文でも出現する割合を計算する方法である
            - METEORを用いる
                - 入力分sと訂正文hに対する意味保存性のスコアS_M(h,s)を次式により求める

$$
\begin{aligned}
P &= \frac{m(h_c)}{|h_c|} \\
R &= \frac{m(s_c)}{|s_c|} \\
\mathrm{S_M(h,s)} &= \frac{P \cdot R}{t \cdot P + (1-t) \cdot R}
\end{aligned}
$$

        - 訂正の前後で文意が変わっていないかという観点
    - ある入力文sに対する訂正文hがであったとき(s,h)に対するスコアを、文法性のスコアS_G、流暢性のスコアS_F、意味保存性のスコアS_Mの重み付き和によって求める

$$
\mathrm{Score}(h,s) = \alpha \mathrm{S_G}(h) + \beta \mathrm{S_F}(h) + \gamma \mathrm{S_M}(h,s),
$$

        - h_cはシステムの出力中の内容語、s_cは原文中の内容語である。m(h_c)は出力中の内容語のうちマッチングされた単語数、m(s_c)は原文中の内容語でマッチングされた単語数を表す。tの値はデフォルト値の0.85
- 実験

- 人手評価との近さ
  - ピアソンの相関係数
  - スピアマンの相関係数
  - 結果

評価尺度	Spearman's ρ	Pearson's r
M^2	0.648	0.632
I-measure	0.769	0.739
GLEU	0.857	0.843
文法性	0.835	0.759
意味保存性	-0.192	0.198
流暢性	0.819	0.864
文法性+意味保存性	0.786	0.771
意味保存性+流暢性	0.929	0.890
流暢性+文法性	0.863	0.844
MFG 中心	0.885	0.878
MFG 最大	0.912	0.898
MFG 最小	0.851	0.854

Table 1: 自動評価による訂正システムのランキングと人手評価間の相関係数.

文法性の必要性が感じられなかった(0.929)

- 文単位評価の性能評価
  - 文単位のスコアを見たとき、自動評価による優劣判定が人手評価と異なっている文があれば、その自動評価尺度は文単位では訂正文を正しく評価できていないことになる
  - 人手評価が異なる２文

原文
On the other hand, the viewers, are not the listeners.

リファレンス
On the other hand, the viewers are not the listeners.

訂正文 A	On the other hand, the viewer, is not the listener.				
	人手評価	提案手法	M^2	I-measure	GLEU
	3	0.892	0.00	-0.391	0.414

訂正文 B	On the other hand, viewer *are* not listeners.				
	人手評価	提案手法	M^2	I-measure	GLEU
	2	0.651	0.714	-0.096	0.496

  - 人手評価が同じ２文

<table>
<tr><td colspan="6">原文</td></tr>
<tr><td colspan="6">With the improvements of technology, a new life with genetic risk can be detected.</td></tr>
<tr><td colspan="6">リファレンス</td></tr>
<tr><td colspan="6">With the improvements in technology, a new life with genetic risk can be detected.</td></tr>
<tr><td rowspan="3">訂正文 A</td><td colspan="5">With the improvement of technology, a new life with genetic risk can be detected.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>5</td><td>0.884</td><td>0.0</td><td>-0.114</td><td>0.449</td></tr>
<tr><td rowspan="3">訂正文 B</td><td colspan="5">With the improvements in technology, a new life with genetic risk can be detected.</td></tr>
<tr><td>人手評価</td><td>提案手法</td><td>M^2</td><td>I-measure</td><td>GLEU</td></tr>
<tr><td>5</td><td>0.874</td><td>1.0</td><td>1.0</td><td>0.566</td></tr>
</table>

- アンサンブルシステム

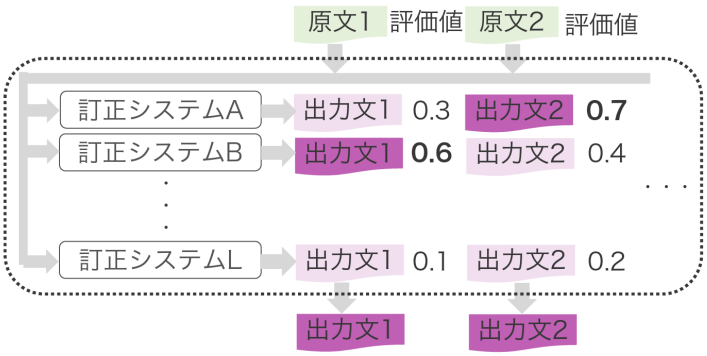

Figure 5: アンサンブルシステムの概要．各システムの訂正を参照無し手法によって評価し，最善の文を出力する．

評価尺度	アンサンブル	トップシステム
TrueSkill	0.462	0.191（AMU）
M^2	0.412	0.372（CAMB）
GLEU	0.548	0.531（CAMB）

Table 8: 訂正システムに対するスコア．トップシステムは CoNLL2014 参加システムで各スコアが最良のシステムを意味し，括弧内にシステム名を示した．

- ニューラル文法誤り訂正モデルによる反復訂正
  - 誤り箇所が多く、文の末尾側の誤りに起因して一回では全ての誤りを訂正するのが困難な文を、繰り返し処理により、徐々に改善できることが期待できる。

The social network *plays* a role in providing information.

主語の数を訂正

Social networks ***plays*** a role in providing information.

動詞を主語と一致させる

Social networks **play** a role in providing information.

- 反復してもあまり改善は見られなかった

Table 9: 訂正の反復の結果

	CoNLL($F_{0.5}$)	JFLEG (GLEU)
無編集	0.0	40.54
1 回目	45.70	51.19
2 回目	46.11	51.79
3 回目	46.19	51.80
4 回目	46.24	51.81
5 回目	46.24	51.81

- 参考文献
  - [1] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. There's no com-
parison: Reference-less evaluation metrics in grammatical error correction. In Proceedings of the 2016 Conference on Empirical Methods in Natural Lan-
guage Processing, pp. 2109-2115. Association for Computational Linguistics,
2016.