ニューラル文法誤り訂正システムにおける リランキングの改善に向けたオラクル分析

小林 正宗 [1]　高橋 悠進 [2]　三田 雅人 [2,3]　小町 守 [2]
[1] 芝浦工業大学　[2] 東京都立大学　[3] 理化学研究所
dz18636@shibaura-it.ac.jp，takahashi-yujin@ed.tmu.ac.jp，
masato.mita@riken.jp，komachi@tmu.ac.jp

概要

ニューラル文法誤り訂正システムは通常，上位 N 文 (N-best) の中からビームサーチを行い，スコアの最も良い文 (1-best) を最終的な出力文として選択する．しかし，1-best はシステム性能の観点から必ずしも最良な文とは限らず，システム性能を最大にする文（オラクル文）を選択することは難しい．この問題の解決策として，特定の情報を利用して N-best の各文のスコアを再計算し並び替えるリランキングがあるが，ニューラル文法誤り訂正においてどのようなリランキングが適切かは明らかになっていない．そこで本研究では，リランキングの改善に向けたオラクル分析を行う．具体的には，オラクル文を上位にリランキングするためにどのような情報が必要か，といった観点でアノテーションを行い，リランキングを改善させる可能性のある要素を分析する．

1　はじめに

文法誤り訂正は，文法的な誤りを含む文を入力とし，誤りを含まない文へと変換し出力するタスクである．近年は，その構造的類似性から，文法誤り訂正タスクを誤りを含む文から誤りを含まない文への翻訳とみなすニューラル機械翻訳に基づくアプローチが主流である．ニューラル文法誤り訂正システムは通常，出力のうち上位 N 文 (N-best) のビームサーチを行い，スコアの最も良い文 (1-best) を最終的な出力文として選択する．しかし，対数尤度の観点からの 1-best は必ずしも文法誤り訂正の観点からの最良な文とは限らず，N-best の中から文法誤り訂正の精度を最大にする文を選択することは難しい．

上記のような問題の解決策として，リランキングが注目されている．リランキングは，N-best の各文のスコアを再計算し並び替える手法であり，元々のシステムに組み込むことが難しい情報や言語資源を利用することができる．リランキングは様々な手法がこれまでに提案されており，Mizumoto ら [1] の統語情報を利用した手法や Chollampatt ら [2] の編集操作と言語モデルの特徴量を利用した手法などがある．リランキングにより，N-best の中から文法誤り訂正精度の高い文を選択することで，文法誤り訂正システムの性能改善が期待できる．

しかし，ニューラル文法誤り訂正においてどのようなリランキングが適切かは明らかになっていない．ここで，**N-best の中からオラクル文を上位にリランキングするにはどのような情報が必要か**，といった問いが自然に提起される．ここでのオラクル文とは，N-best の中から任意の評価手法において精度が最も高くなるような文を指し，そのときの精度はそのシステムの理論上の上限値を意味する．また，この問いについて調査することは，単にリラン

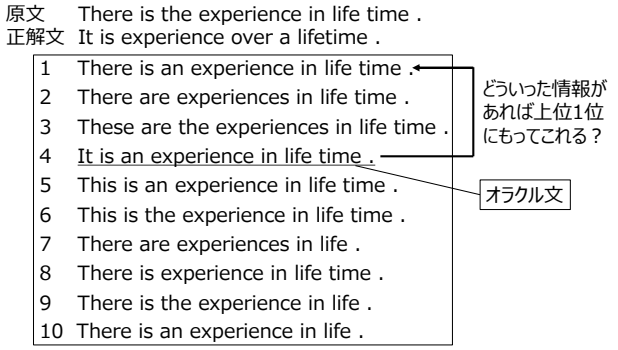

図 1　本研究の概要

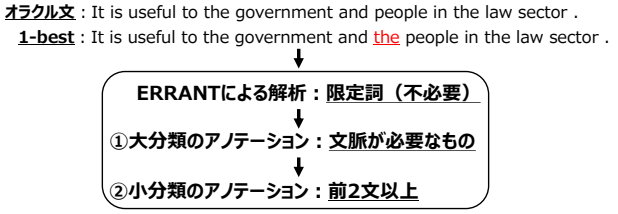

図 2　アノテーションの概要

表 1 アノテーションの分類ラベルと例．**太字**は誤り箇所，青色のハイライトはアノテーション時の判断根拠となる箇所をそれぞれ示している．

大分類	小分類	例文
文内で解ける	（エラータイプ）	
	動詞の時制	This technology can also *be* a helpful tool to assist law enforcers such as the police to collect evidence of crime after **[*it has occurred → they occur]** .
	限定詞	Additionally , the service industry were also using it to track **[*the → its]** *clothes* that were being used by their employees .
	名詞の単複	This can increase the efficiency of *solving* **[*crime → crimes]** and thus decrease the amount of criminal acts committed by people .
文外の文脈が必要	（文脈幅）	
	前 1 文	In addition , if surveillance technology is even better developed , it can be used to detect **[*the problem → problems]** before the real accident has happened .
	前 2 文以上	Firstly , security systems **[*are → have]** improved in many areas such as school campuses or at the workplace .
	後 1 文	We should respect that everyone has human rights and without **[*others ' → other 's]** permission , this should not be placed or implanted on anyone .
	後 2 文以上	No matter how good and how many security systems are there for us , **[*hackers → the hackers]** are also increasingly improving their skills in terms of hacking and breaking the codes of the security system .
	解けないもの	However , I think such a powerful **[*device → devices]** shall not be made easily available to the public or fall into the hands of 'ill-intentioned ' individuals .
一般的な知識が必要		The convenience and high efficiency of using electronic products **[*is being → has been]** noticed by people worldwide .
誤りではない		In the modern digital world , electronic products are widely used in daily lives such as smart phones , computers **[, → and]** etc .

キング手法を改善する可能性のある要素を明らかにするだけでなく，最先端システムを対象に調査することで，現状のシステムにおける訂正が困難な文法誤りを明らかにすることに繋がると考えられる．

そこで本研究では，上記の問いの答えを調査すべく，リランキングの改善に向けたオラクル分析を行う．本研究のイメージを図 1 に示す．具体的には，まずはじめに最先端文法誤り訂正システムを対象として，オラクルを算出し，システムの 1-best とオラクル文には大きな差があることを定量的に示す．次に，オラクル文を上位にリランキングするにはどのような情報が必要か，といった観点で人手でアノテーションを行い，リランキングを改善させる可能性のある要素を明らかにする．

本研究の主要な貢献は以下である。

- 分析対象データのうち，8 割がリランキングで改善する余地があることがわかった．5 割は文内で解けるものであり，文内の文脈の重要性が示された．残りの 3 割は文外の文脈が必要なものであり，特に前の数文を考慮することは有効であると示された．
- 分析対象データのうち，2 割がリランキングでの改善が難しいことがわかった．リランキング手法のみならず，評価手法についても改善の余地がある可能性が示された．

2 分析手法

本研究では，オラクル文を上位にリランキングするために必要となりうる情報を得るために，オラクル文を正解文とみなしニューラル文法誤り訂正システムの 1-best 出力に対してアノテーションを行う．具体的なアノテーションの概要を図 2 に示す．

まず，オラクル文を正解文，1-best を原文とみなし，文法誤り訂正システムの評価・分析ツールキット ERRANT [3] を用いて 1-best 文の誤り箇所，エラータイプ，編集操作を解析する．次に，ERRANT の解析結果をもとに，1-best に対してアノテーションを行う．アノテーションは基本的に文を見て行う．また，アノテーションをする際は文内の文脈で決まらない時のみ文外の文脈を使用する．

アノテーションの分類ラベルは大分類と小分類の 2 つに分ける．大分類と小分類の概要およびそれぞれのアノテーションの具体例を表 1 に示す．大分類の種類は，**文内で解ける**，**文外の文脈が必要**，**一**

般的な知識が必要，**誤りではない**の計 4 つである．大分類は，1-best をオラクル文にするための訂正内容を明示するラベルであり，オラクル文に近づけるために最尤なラベルを 1 つアノテーションする．小分類は，大分類を細分化したラベルであり，文内で解けるものは ERRANT によるエラータイプ，文外の文脈が必要なものは文脈幅で細分化を行った．また，一般的な知識が必要なものと誤りではないものは細分化は行わなかった．小分類は文内の誤り箇所ごとにアノテーションする．

3 実験設定

ベースシステム　実験に使用する文法誤り訂正システムとして，Kiyono ら [4] のシステムを使用する．Kiyono らのシステムは文法誤り訂正の代表的なベンチマークである CoNLL2014 [5]，および BEA-2019 [6] において世界最高水準の性能を達成しており，文法誤り訂正の現状を分析をするうえで適していると考えられる．具体的には，本研究では Kiyono らのシステムにおける PRETLARGE+SEE (finetuned) モデルを使用した [1]．なお，ハイパーパラメータは，バッチサイズを 16，N-best の N に相当するビーム幅を 10 とし，それ以外を全て Kiyono らと同じ値に設定した．

データ　本実験では，CoNLL-2013 [7] を分析対象のデータとして使用する．CoNLL-2013 テストセットは，CoNLL-2013 共通タスクで提供されたデータセットであり，文法誤り訂正分野では，開発データとして一般的に使用されている．本実験では，アノテーションの前に，CoNLL2013 テストセットの全 1,381 文の中から 1-best とオラクルが異なるようにサンプリングした 70 文に対し，それぞれ ERRANT を用いて 1-best の誤り箇所，エラータイプ，編集操作の解析を行った．ここでの N-best は，CoNLL2013 テストセットを Kiyono らのシステムに入力した出力のうち，出力確率の高い上位 10 文である．オラクル文は，N-best の各文に対して M2Scorer [8] で $F_{0.5}$ を計算した中で最も $F_{0.5}$ スコアが高い文である．そして，サンプリングした 70 文に対し，大分類，小分類の順でアノテーションを行った．アノテーションは 2 名で行った．

1）実装および事前学習済みモデルは，https://github.com/butsugiri/gec-pseudodata にて著者らが公開しているものを使用した

表 2　ERRANT によるエラータイプ（上位 5 項目）

	項目	#	%
エラータイプ	限定詞	27	24
	動詞の時制	18	16
	前置詞	12	11
	名詞の単複	10	9
	代名詞	6	5
	計	111	100

表 3　大分類のアノテーション結果

項目	#	%
文内で解ける	37	53
文外の文脈が必要	18	26
一般的な知識が必要	1	1
誤りではない	14	20
計	70	100

4 実験結果

表 2 に ERRANT による 1-best の解析結果を示す．エラータイプのうち，限定詞の誤りが最も多く，次点で動詞の時制，前置詞の誤りが多かった．

表 3 に大分類のアノテーション結果を示す．アノテーションの妥当性を測るために，2 名のアノテーターの agreement を測ったところ，62.9%であった．Cohen's kappa の値は 0.39 で，Landis and Koch [9] の基準では fair agreement に該当した．

表 4 に小分類のアノテーション結果を示す．文内で解けるもののうち，動詞の時制の誤りが最も多く，次点で前置詞，限定詞の誤りが多かった．文外の文脈が必要なもののうち，39%（33 箇所中の 13 箇所）は実際に使用しても解けないものであった．また，文外の文脈を使用して解けるもののうち，85%（20 箇所中の 17 箇所）は前の文外の文脈を使用することで解けることがわかった．

5 分析・考察

本節では，リランキングを改善させる可能性のある要素を明らかにするために，大分類の 4 つの分類ラベルごとに 1-best の分析を行う．

文内で解けるもの　文内で解ける 37 文（表 3）とサンプリングした 70 文を，ERRANT によるエラータイプ上位 5 項目で比較をした．その結果，文内における 37 文における限定詞の割合が 15%減少して

表 4 小分類のアノテーション結果

	項目	#	%
文内で解ける	動詞の時制	11	20
	前置詞	8	15
	限定詞	5	9
	句読点	5	9
	名詞の単複	4	7
	計	55	100
文外の文脈が必要	前 1 文	9	27
	前 2 文以前	8	24
	後 1 文	2	6.1
	後 2 文以降	1	3.0
	解けないもの	13	39
	計	33	100
一般的な知識が必要		2	100
誤りではない		21	100

いた．そのため，文内の情報では，限定詞の誤りの訂正が難しいことが考えられる．

文外の文脈が必要なもの　文外の文脈が必要と考えられる誤り箇所の 61%（33 箇所中の 20 箇所）が実際に前後の数文を参照することで解くことができた．また，後ろの数文よりも前 1 文または前 2 文以前を参照すると解けるものが多いことがわかった．従って，リランキングの際に前後の出力文を考慮することは有効であり，特に前の数文を考慮することはリランキング性能を大きく向上させる可能性が考えられる．

一般的な知識が必要なもの　一般的な知識が必要なものはサンプリングした 70 文のうち 1 文であったことから，常識やイディオムなどの一般的な知識はリランキング性能を向上させる上で必要な要素ではあるが，比較的その重要性は低いことがわかった．

誤りではないもの　誤りではないものは 20%存在していた．これは，オラクル文に対して 1-best が正当に評価されていないことを示している．つまり，1-best とオラクル文の文意の差が大きくないにも関わらず，1-best を過小評価してしまっている可能性がある．従って，リランキング手法のみならず，文法誤り訂正システムの評価手法についても改善の余地があると考えられる．

6　関連研究

文法誤り訂正システムの性能向上のために，様々な N-best のリランキング手法の研究が行われている．Mizumoto らはフレーズベース統計的機械翻訳を使い，品詞や依存関係などの統語情報を活用することでリランキング性能を向上した．Chollampatt らは対数線形モデルを使い，システムの候補文のスコアに編集操作と言語モデルによる特徴量を追加し，スコアの再計算をする手法を提案した．Hoang ら [10] は統計的機械翻訳システムによる文の修正を有効または無効に分類する分類器を構築し，N-best に分類器を使うことで各候補文を分類し，スコアを再計算することでリランキングを行った．Liu ら [11] は複数の候補文を用いて文法誤り訂正の品質推定を行う VERNet を提案し，トークンレベルの品質推定スコアを用いて高性能のリランキングを実現した．

7　おわりに

本研究では，リランキング性能を改善させる可能性のある要素を分析するために，オラクル文を上位にリランキングするためにはどのような情報が必要かという観点で 1-best に対してアノテーションを行った．アノテーションの結果より，リランキングにおいて文内の情報を利用することは，動詞の時制や限定詞，代名詞などの長距離の文脈を必要とする際に有効であると考えられる．また，文外の文脈を考慮することが性能を向上させる有効な要素となりうると判明した．さらに，リランキング手法のみならず，評価手法についても改善の余地がある可能性が示された．

今後は，本研究で得られた知見を利用することによる，文内および文外の文脈を利用したリランキング性能に関する研究を行いたい．

参考文献

[1] Tomoya Mizumoto and Yuji Matsumoto. Discriminative reranking for grammatical error correction with statistical machine translation. In **NAACL**, 2016.

[2] Shamil Chollampatt and Hwee Tou Ng. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In **AAAI**, 2018.

[3] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **ACL**, 2017.

[4] Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In

EMNLP, 2019.

[5] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In **CoNLL**, 2014.

[6] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In **BEA**, 2019.

[7] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 shared task on grammatical error correction. In **CoNLL**, 2013.

[8] Daniel Dahlmeier and Hwee Tou Ng. Better evaluation for grammatical error correction. In **ACL**, 2012.

[9] J. Richard Landis and Gray G. Koch. The measurement of observer agreement for categorical data. In **Biometrics**, 1977.

[10] Duc Tam Hoang, Shamil Chollampatt, and Hwee Tou Ng. Exploiting n-best hypotheses to improve an SMT approach to grammatical error correction. In **IJCAI**, 2016.

[11] Zhenghao Liu, Xiaoyuan Yi, Maosong Sun, Liner Yang, and Tat-Seng Chua. Neural quality estimation with multiple hypotheses for grammatical error correction. In **NAACL**, 2021.