リランキングによる文法誤り訂正/訂正候補提示の性能改善

水本智也
東北大学
tomoya-m@ecei.tohoku.ac.jp

松本裕治
奈良先端科学技術大学院大学
matsu@is.naist.jp

1 はじめに

第二言語学習を支援する研究が注目されており，自然言語処理分野においては文法誤りの訂正の研究が多く行なわれている．英語の文法誤り訂正の性能を競うコンペティションも 2011 年から 4 年連続で開催された．

全ての誤りタイプを扱うための手法として，統計的機械翻訳 (SMT) を使った文法誤り訂正手法が提案されている [7, 4]．SMT を使った手法は CoNLL2014 Shared Task において 1 番と 3 番の性能を達成した [10]．

SMT の手法では，候補文を複数作成し，その候補に対してスコア付けを行ない，もっともスコアの高いものを翻訳文として出力する．このスコア付けの問題は難しく，候補中でもっともよい翻訳が 1 番高いスコアにならないことがある．言い換えると，他の候補の中に最もよい翻訳が含まれていることがある．同様の問題は SMT を用いた文法誤り訂正でも生じる．

このスコア付けの問題を解決する方法のひとつにリランキングがある．リランキングは，SMT システムが出力した上位 N 個（の候補以下，N-best）を再びスコア付けしなおして並べ替える手法であり，一般的な機械翻訳タスクにおいてリランキング手法が提案されている [1]．図 1 にリランキングの流れを示す．最初に SMT を使った文法誤り訂正システムに学習者の文を入力して N-best を得る（図 1 中の青の破線 [A]）．次にリランキングシステムは N-best に対して新たにスコア付けを行ない，訂正候補の並び替えを行なう（図 1 中の赤の破線 [B]）．

先行研究ではリランキングの手法は 1-best 出力の結果を改善するために用いられてきたが，リランキングは文法誤り訂正の性能改善だけでなく，訂正候補提示の観点から見ても効果が期待できる．学習者支援を考えた場合，システムが 1 つ訂正を出力するよりも，いくつか訂正候補として提示を行ない，学習者自身で選択できる方がよいと言える．実際，提示された候補の質を学習者が見分けられると報告されている [5]．このとき，上位によい訂正候補があれば，また上位に悪い訂正候補がなければ訂正候補提示としてはよいシステムと言える．リランキングを行なうことで，よい訂正候補を上位に，悪い訂正候補を下位にすることができると考える．

本稿では，リランキングの手法を英語の文法誤り訂正に適応し，訂正候補提示の観点からの評価も行なう．文法誤り訂正で使われているフレーズベース SMT では品詞や係り受けといった情報は考慮できない．本稿では品詞や係り受けといった情報を使用しリランキングを行ない，リランキングによる文法誤り訂正，訂正候補提示においてリランキング，品詞や係り受けの情報が有効であることを示す．

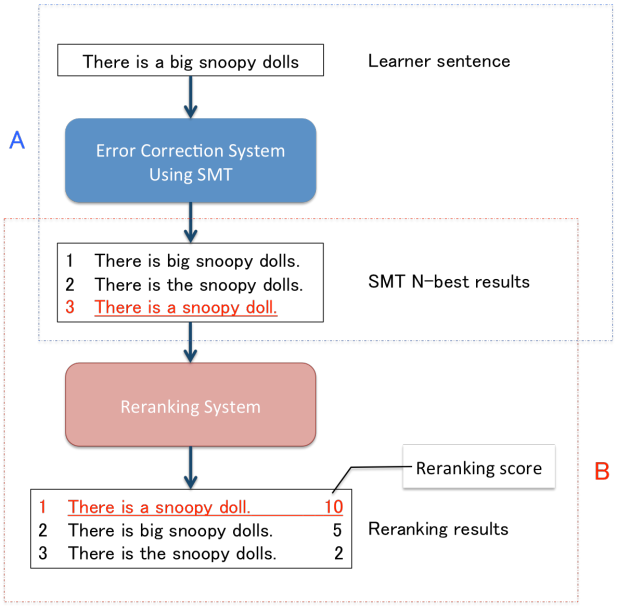

図 1 リランキングの流れ

2 関連研究

2.1 リランキングに関する関連研究

機械翻訳タスクに対するリランキングの手法が提案されている [13, 1, 6]．Shen ら [13] は最初に機械翻訳タスクに対してリランキングを適用した．彼らの研究では少ない素性を使ってリランキングを行なった．Li ら [6] は機械翻訳タスクにおいて，大規模な識別的 N-gram 言語モデルを使ってリランキングを行なった．Carter ら [1] の手法は Li ら [6] に似ているが，彼らは統語的な情報を使ってリランキングを行なった．

フレーズベース SMT の手法を使った文法誤り訂正においてもリランキングの手法は提案されている [4]．彼らの手法は大規模な言語モデルを用いてリランキングを行なう．しかしながら，リランキングステップでは表層情報だけでなく，Carter ら [1] が使ったような統語的な

情報も使うことができる．我々のリランキングシステムでは統語情報を考慮する．

2.2 訂正候補提示に関する関連研究

訂正候補提示を行なっている研究は英語の動詞の語彙選択を対象にしたもの [12] と日本語のコロケーションを対象としたもの [11] がある．これらのタスクでは，候補提示するための候補を探すことが難しいタスクであり，彼らの研究では，言語学習 SNS Lang-8 のデータを使って訂正候補を大規模に獲得することでこの問題を解決している．

我々の研究では，先行研究と違い誤りのタイプを限定しない訂正候補提示を行なう．さまざまな誤りタイプを含む文のリランキングをするため，先行研究と比べて難しいと言える．一方で訂正候補自体は SMT の誤り訂正システムの N-best の中のものを使用するため，先行研究のように訂正候補の獲得する必要はない．

3 リランキングは必要か？

SMT を用いた文法誤り訂正も機械翻訳タスクと同様の問題があり，文法誤り訂正システムの 1-best の訂正がもっともよい訂正であるとは限らない．これを確かめるために，SMT を使った文法誤り訂正を行ない，N-best 中のオラクルスコアを計算する．オラクルスコアは学習者文に対する文法誤り訂正システムの出力の N-best の中から，もっともスコアが高くなる訂正を選んだ場合のスコアである．

表 1 に SMT を用いた文法誤り訂正のベースラインシステムのオラクルスコアを示す[*1]．システムの 1-best 出力の $F_{0.5}$ は 37.9 であるのに対して，10 個の候補を出した場合のオラクルの $F_{0.5}$ は 64.3 である．N-best の N の値が大きくなるにつれて，オラクルスコアも高くなる．この結果から SMT による文法誤り訂正の 1-best の訂正がもっともよい訂正ではないことがわかる．

■文法誤り訂正におけるリランキングの利点　文法誤り訂正においてリランキングを行なう利点は 3 つある．1 つ目の利点は，最初に用いる SMT システム（本稿ではフレーズベース SMT）で扱うことのできない品詞や統語的な情報をリランキングの際に利用できることである．誤りの中には離れた単語間の関係を考慮する必要のあるものがある；例えば，*a big Snoopy dolls* にある *a* と *dolls* のような関係がある．そのため，統語的な情報は文法誤り訂正に有効であると考える．

2 つ目の利点は品詞タガーや係り受け解析器が，誤りの訂正された候補文に対して解析を行なうことができる点である．そのため誤りの含まれる文を解析するよりも頑健に解析でき，リランキングの際により正確な素性として扱うことができる．

[*1] ベースラインシステムについては 5.1 で説明する．

表 1　文法誤り訂正のオラクルスコア

N-best	Precision	Recall	$F_{0.5}$
1	43.9	24.5	37.9
10	79.1	36.7	64.3
50	89.5	43.1	73.6
100	92.3	45.3	76.4

3 つ目の利点は，リランキングすることで誤り訂正だけでなく，訂正候補提示としてのシステムの性能改善が期待できる点である．訂正候補提示を考えると，システムの 1-best に正解の訂正が来なくても，なるべく上位に正しい訂正が出現するとよいと言える．ベースラインシステムでは下に出現する正しい訂正を，リランキングを行なうことで上位にもってくることが可能であると考える．

4 リランキング手法

4.1 識別的リランキングアルゴリズム

本稿では，SMT のタスクで使用されたパーセプトロンを用いたリランキングを行なう．図 2 にパーセプトロンを使ったリランキングのアルゴリズムを示す．T はパーセプトロンの学習のイテレーション数であり，N はトレーニングコーパス中に含まれる文数である．$GEN(x)$ は入力文に対して，SMT を使った文法誤り訂正システムによって生成された N-best である．$ORACLE(x^i)$ は N-best の中で $F_{0.5}$ が最も高くなる文とする．w は素性に対する重みベクトルであり，ϕ は各候補文に対する素性ベクトルである．各候補文に対してスコアを計算して，オラクルの文と異なる場合に重みを更新する．本稿では候補文として，人手で添削されたゴールドデータも含めて学習を行なう．これは，正解の文によく出てくる素性が高い重みを持つように学習するためである．重みの学習には平均化パーセプトロンを用いた．

N ベストリストから最もよい候補を選ぶ計算式として以下を用いる．

$$S(z) = \beta\phi_0(z) + w \cdot \phi(z)$$

$\phi_0(z)$ は，SMT で誤り訂正をした際のスコアであり，β によって重み付けする．SMT のスコアをリランキングの素性として使用すると under-training を引き起こすため，このように候補選択のための最終的なスコアを計算する．β の値は開発データによって決定する．

4.2 リランキングのための素性

本稿では，Carter らの用いた素性 [1] に加えて，新たにいくつかの素性を用いる．Carter らの素性は表層単語系列，POS tag 系列，shallow-parse tag 系列および POS-tag と shallow-parse tag を合わせた系列から生成される．

新たに追加した素性を表 2 に示す（以下，これらの素性を New Feature と呼ぶ．）．品詞-表層 2,3,4,5-gram 素性は，内容語は品詞に置換し，機能語は表層形のまま表現する N-gram 素性である．Web 係り受け N-gram 素性

表 2　リランキングのための素性．例は文 *I agree with this statement to a large extent* に対する素性を示す．“Web dependency N-gram” 以外はバイナリ素性として表現し，“Web dependency N-gram” は 0〜1 の実数値素性である．

素性名	例
単語表層 2,3-gram	I agree; I agree with; agree with; agree with this; this statement
品詞 2,3,4,5-gram	PRP VBP; PRP VBP IN; PRP VBP IN DT; PRP VBP IN DT NN
品詞-表層 2,3,4,5-gram	PRP VBP; PRP VBP with; PRP VBP with this; PRP VBP with this NN
Web 係り受け N-gram	prep-agree-with-statment; det-a-extent

1: $w \leftarrow 0$
2: **for** $t = 1$ to T **do**
3: **for** $i = 1$ to N **do**
4: $y^i \leftarrow ORACLE(x^i)$
5: $z^i \leftarrow argmax_{x \in GEN(x^i)} \phi(z) \cdot w$
6: **if** $z^i \neq y^i$ **then**
7: $w \leftarrow w + \phi(y^i) - \phi(z^i)$
8: **end if**
9: **end for**
10: **end for**
11: **return** w

図 2　パーセプトロンを使ったリランキングアルゴリズム

は，[2] が冠詞，前置詞を訂正するために用いた素性である．大規模な係り受けの付いたコーパスから係り受け N-gram を取り出して頻度をカウントしログを取り，その値が 0〜1 の間の実数になるように正規化する．

5　実験

SMT を用いた文法誤り訂正および訂正候補提示におけるリランキングの効果を調べるために実験を行なった．

5.1　実験設定

フレーズベース SMT のツールとして，cicada 0.3.5 を使用した．デコーダ，単語アライメントは cicada 0.3.5 の内部実装を用いた．言語モデルには KenLM を使用し，5-gram 言語モデルを構築した．SMT のモデルのパラメータ調整には ZMERT を使用し，$F_{0.5}$ を最適化するようにパラメータのチューニングを行なった．

トレーニングデータとして “Lang-8 Learner Corpora v2.0” を使用した．本稿では Lang-8 Learner Corpora を用い，データに含まれるノイズを除くため文献 [8] の方法を元に挿入，削除ともに 5 以下のもののみ使用した．この結果，1,069,127 文対が抽出され，これを翻訳モデルに使用した．言語モデルは，“English Gigaword” と “The NUS Corpus of Learner English” [3] から構築し，それぞれ別の素性関数として使用した．リランキングモデルの学習は，Lang-8 Learner Corpora を 10 分割し，9 個で SMT のモデルを学習し 1 つに対して訂正を，10 回繰り返すことで作成した．

評価データとチューニングには，それぞれ CoNLL-2014 テストセット，CoNLL-2013 テストセットを使用した．CoNLL-2013 テストセットを 700 文と 681 文に分割し，SMT モデルのチューニングとリランキングの β の値の決定にそれぞれ用いた．CoNLL-2014 テストセットの 1,312 文を評価用コーパスとして使用した．

5.2　評価方法

文法誤り訂正の評価手法として，$F_{0.5}$，GLEU [9] を用いる．$F_{0.5}$ の算出には m2scorer[*2] を使用した．GLEU は，機械翻訳タスクで使用される評価尺度 BLEU を誤り訂正用に改良したものであり，個々の誤りが正解しているかでなく，文全体のスコアを計算することで評価する．

訂正候補提示の評価には平均逆順位 (MRR) を用いる．

$$MRR = \frac{1}{N}\sum_{i=1}^{N} RR_i; RR_i = \begin{cases} \frac{1}{r(gold)} & (gold_i \in S_i) \\ 0 & otherwise \end{cases}$$

式中の N は事例数を表し，ここでは評価コーパス中の文数である．事例 i に対する逆順位 RR_i は，候補リスト S_i 中の正解 $gold_i$ の順位 $r(gold_i)$ によって計算される．本稿の実験では，正解 $gold_i$ は N-best の中でもっとも $F_{0.5}$ が高くなる訂正とする．推薦した候補中に正しい訂正が含まれてない場合は逆順位は 0 になるが，本稿では N-best 全てを使って MRR を計算するため逆順位が 0 になることはない．

5.3　実験結果

実験結果を表 3 に示す．比較のためのベースラインとして，SMT の 1 ベスト出力と，言語モデル確率によるリランキング [4] を用いた．また，CoNLL2014 Shared Task において 1 番目，2 番目であった CAMB チームと CUUI チームの出力とも比較する．New Feature を使ったリランキングシステムが $F_{0.5}$，GLEU，MRR においてベストの値を達成した．

単純に N-gram 言語モデルの確率でリランキングを行なうと，true positive が大きく増えるため，Recall は大きく向上するが，false positive も大きく増えてしまい Precision が下がる．N-gram 言語モデルの結果は CAMB と非常に似た結果となっているがこれは CAMB のシステムが SMT + N-gram 言語モデルによるリランキングを行なっているからであると考える．

単語 2,3-gram を使った識別的リランキングは $F_{0.5}$，GLEU においてベースラインと比べて差がない結果となった．これはベースラインの時点でより大きな言語モデルを使っているためだと考える．Carter らの素性を

[*2] https://github.com/KentonMurray/Non-nativeEnglishGrammarCorrection/tree/master/release2.2/m2scorer

表 3　文法誤り訂正のおよび訂正候補提示の評価結果. TP，FN，FP はそれぞれ true positive，false negative，false positive を表す．アスタリスクはある手法が SMT Baseline と比較して統計的有意差があることを示す ($p < 0.01$).

	システム	Precision	Recall	$F_{0.5}$	TP	FN	FP	GLEU	MRR
	ベースライン								
1	SMT ベースライン	43.9	24.5	37.9	598	1847	764	65.7	60.9
2	N-gram 言語モデル	39.5	**31.7**	37.6	834	1797	1280	64.7	44.4
3	CAMB (CoNLL2014)	39.7	30.1	37.3	772	1793	1172	64.5	-
4	CUUI (CoNLL2014)	41.8	24.9	36.8	623	1881	868	64.8	-
	識別的リランキング								
5	単語 2,3-gram	43.7	24.8	37.9	606	1834	781	65.7	61.3
6	Carter (2011)	44.3	26.7	39.1	669	1837	842	65.8	61.5
7	New Feature (表 2)	**45.8**	26.6	**40.0***	657	1813	778	**66.1**	**63.3***
8	6+7	44.4	27.1	39.4*	679	1827	851	65.8	61.2

使った場合は，$F_{0.5}$（ベースラインと比べて p¡ 0.05 で統計的有意差あり），GLEU において性能改善が見られ POS や shallow-parse 素性が文法誤り訂正でも有効であることがわかる．New Feature を用いたリランキングでは，Carter らの素性よりもさらに性能改善がされており，品詞-表層素性や係り受け素性が有効であることがわかる．Carter らの素性と New Feature 両方を使うと，New Feature のみよりも全体的に低い値となった．これは，shallow-parse 素性と係り受けなど役割が重なっている素性があるためだと考える．MRR による訂正候補提示の評価を見ると，New Feature を用いたリランキングでは MRR が大きく上がっている一方で，他の素性を用いたリランキングではほとんど向上が見られなかった．

各リランキングシステムの結果を文の長さ（単語数）別に分析を行なう．まず文法誤り訂正の結果だが，長さが一桁の文である場合の平均の $F_{0.5}$ を比較すると，New Feature を使ったリランキングシステムが他のリランキングシステムより約 2 ポイント高い．短い文ほど係り受け解析などの失敗が少なく，より係り受け解析素性が効いてるからだと考える．

訂正候補提示について，文の長さ別にベースラインと New Feature のリランキングシステムで比較する．文の長さが一桁の場合の MRR は大きく差がない一方で，文が長くなると New Feature のリランキングシステムの方が MRR が高くなる．特に文の長さが 40 以上になると，New Feature のリランキングシステムの方が 5 ポイント以上高い結果となる．ベースラインシステムでは長い文の訂正が難しく，正しい訂正が上位にこないのに対し，New Feature のリランキングシステムは係り受けなどを使うことで，上位の方に正しい訂正を持ってくることができていると考える．

6　おわりに

本稿では，SMT を用いた英語文法誤り訂正および訂正候補提示に対してリランキングを行なうことで性能改善を行なった．SMT のタスクで提案されたパーセプトロンによるリランキングを英語誤り訂正/訂正候補提示に応用した．最初の SMT システムで使用していない品詞や係り受けといった素性を用いて，リランキングすることで文法誤り訂正，訂正候補提示の両方において性能改善可能であることを示した．今後は，文法誤り訂正に効果的な素性の開発，リランキングの学習アルゴリズムの変更，リランキング自体のアルゴリズムの変更など行なう予定である．

参考文献

[1] S. Carter and C. Monz, “Syntactic Discriminative Language Model Rerankers for Statistical Machine Translation,” Machine Translation, vol.25, no.4, pp.317–339, 2011.

[2] D. Dahlmeier, H.T. Ng, and E.J.F. Ng, “NUS at the HOO 2012 Shared Task,” Proceedings of BEA, pp.216–224, 2012.

[3] D. Dahlmeier, H.T. Ng, and S.M. Wu, “Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English,” Proceedings of BEA, pp.22–31, 2013.

[4] M. Felice, Z. Yuan, Ø.E. Andersen, H. Yannakoudakis, and E. Kochmar, “Grammatical error correction using hybrid systems and type filtering,” Proceedings of CoNLL Shared Task, pp.15–24, 2014.

[5] C. Leacock, M. Gamon, and C. Brockett, “User Input and Interactions on Microsoft Research ESL Assistant,” Proceedings of BEA, pp.73–81, 2009.

[6] Z. Li and S. Khudanpur, “Large-scale Discriminative n-gram Language Models for Statistical Machine Translation,” Proceedings of AMTA, 2008.

[7] T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto, “The Effect of Learner Corpus Size in Grammatical Error Correction of ESL Writings,” Proceedings of COLING, pp.863–872, 2012.

[8] T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto, “Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners,” Proceedings of IJCNLP, pp.147–155, 2011.

[9] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, “Ground Truth for Grammatical Error Correction Metrics,” Proceedings of ACL-IJCNLP, pp.588–593, 2015.

[10] H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant, “The CoNLL-2014 Shared Task on Grammatical Error Correction,” Proceedings of CoNLL Shared Task, pp.1–14, 2014.

[11] L. Pereira, E. Manguilimotan, and Y. Matsumoto, “Automated Collocation Suggestion for Japanese Second Language Learners,” Proceedings of ACL Student Research Workshop, pp.52–58, 2013.

[12] Y. Sawai, M. Komachi, and Y. Matsumoto, “A Learner Corpus-based Approach to Verb Suggestion for ESL,” Proceedings of ACL, pp.708–713, 2013.

[13] L. Shen, A. Sarkar, and F. Josef Och, “Discriminative Reranking for Machine Translation,” Proceedings of HLT-NAACL, 2004.