6C-08

統計的機械翻訳を用いた英語文法誤り訂正のリランキングによる性能改善

水本智也　　松本裕治
奈良先端科学技術大学院大学
{tomoya-m, matsu}@is.naist.jp

1　はじめに

英語学習者の文法誤りの訂正を行なう研究が盛んになっており，英語の文法誤り訂正の性能を競うコンペティションも4年連続で開催されている．文法誤り訂正の手法として統計的機械翻訳を用いる手法が提案されており [5]，上記のコンペティションでも高い性能を出している．

統計的機械翻訳の手法では，候補の文を複数作成し，その候補に対してスコア付けを行ない最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，候補中で最も良い翻訳が1番高いスコアにならないことがある．同様の問題が文法誤り訂正に統計的機械翻訳を用いた場合にも生じる．表1に統計的機械翻訳を用いて文法誤り訂正を行なったオラクルスコアを示す[*1]．オラクルスコアはNベスト出力の中で最も良い訂正を選ぶことができた場合のスコアである．1ベストのみ出力した場合の $F_{0.5}$ が40.96，10ベストを出力した場合のオラクルスコアの $F_{0.5}$ が68.95となり，出力数を増やすことでスコアが上昇している．

この問題を解決する手法のひとつとしてリランキングがある．リランキングは，システムが出力したNベストの結果を再びスコア付けして並べ替える手法である．一般的な統計的機械翻訳のタスクにおいてリランキング手法が提案されている [1]．文献 [1] のリランキングでは，最初の翻訳システムで考慮していない品詞情報や構文情報を組み込むことで性能改善を行なっている．

文法誤り訂正の研究でもリランキングを行なっている研究がある [4]．文献 [4] は統計的機械翻訳を用いて誤り訂正を行なった後，言語モデルの確率を用いてリランキングを行なっている．文献 [4] のリランキングでは，文献 [1] で考慮した品詞情報や構文情報を使っておらず，統計的機械翻訳でも考慮している言語モデルのみでリランキングしている．そこで本稿では，文献 [1] で用いた手法を文法誤り訂正に応用してリランキングし，品詞，構文情報を考慮することで誤り訂正の性能向上できることを示す．

表1　統計的機械翻訳を用いた英語文法誤り訂正のオラクルスコア

N-best	Precision	Recall	$F_{0.5}$
1	50.96	22.18	40.46
10	87.61	37.23	68.95
30	92.32	42.57	74.83
50	94.33	44.57	77.11
100	95.95	47.22	79.53

2　リランキング手法

本稿では，統計的機械翻訳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1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．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，人手で添削し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．

$$S(z) = \beta\phi_0(z) + w \cdot \phi(z) \tag{1}$$

$\phi_0(z)$ は，統計的機械翻訳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．

リランキングに用いる素性を表2に示す．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N-gram素性である．係り受け素性は，文法誤り訂正に有効であると考えるものを使用する[*2]．

[*1] 用いたデータ等の詳細は3節で示す

[*2] 冠詞-名詞ペア，主語-動詞ペアなど

表 2　文 “I agree with this statement to a large extent .” に対する素性の一例

ID	素性	例
1	単語 2,3-gram	I agree; I agree with; agree with; agree with this
2	品詞 2,3,4,5-gram	PRP VBP; PRP VBP IN; PRP VBP IN DT; PRP VBP IN DT NN
3	品詞-表層 2,3,4,5-gram	PRP VBP; PRP VBP with; PRP VBP with this; PRP VBP with this NN
4	単語/品詞ペア	I/PRP; agree/VBP; with/IN
5	係り受け	nsubj(agree, I); det(statement, this); prep_with(agree, statement)

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 1　パーセプトロンを使ったリランキングアルゴリズム

3　実験

統計的機械翻訳を用いた文法誤り訂正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．

フレーズベース統計的機械翻訳のツールとして，cicada 0.3.5 を使用した．単語アライメントも 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．統計的機械翻訳のモデルのパラメータ調整には ZMERT を使用し，F 値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．評価には m2scorer [2] を使用した．

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 [6] の方法を元に挿入，削除ともに 10 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用いた．評価データとチューニングには，それぞれ CoNLL2014 Shared Task，CoNLL2013 Shared Task のデータを使用した．

リランキングモデルの学習データは，Lang-8 Learner Corpora で 10 分割し，9 個で統計的機械翻訳のモデルを学習し 1 つに対して訂正を行なうことを 10 回繰り返して作成した．β の値はチューニングデータを用いて決定した．

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，統計的機械翻訳の 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．単純に言語モデル確率でリランキングを行なうと，Recall は大きく向上するが，Precision が下がる．パーセプトロンを用いたリランキングでは，単語/品詞素性 (4) を用いた場合を除いて，Precision をほぼ下げることなく Recall を向上させることができ，結果として $F_{0.5}$ の向上に成功した．

表 3　実験結果

	Precision	Recall	$F_{0.5}$
Baseline (Original)	50.96	22.18	40.46
Baseline (LM Reranking)	39.45	**31.70**	37.61
Our Reranking（素性 1）	50.95	22.23	40.54
Our Reranking（素性 2）	49.71	24.35	41.13
Our Reranking（素性 3）	50.82	23.22	41.06
Our Reranking（素性 4）	48.74	23.96	40.38
Our Reranking（素性 5）	**51.19**	22.38	40.71
Our Reranking（素性 ALL）	50.67	24.67	**41.85**

4　おわりに

本稿では，統計的機械翻訳を用いた英語誤り訂正のリランキングを行なった．統計的機械翻訳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に応用した．素性は単純なもののみを使用したが，リランキングすることで性能改善可能であることを示した．今後は，誤り訂正により効果的な素性の開発を行なう予定である．

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 and H.T. Ng, “Better evaluation for grammatical error correction,” Proceedings of NAACL-HLT, pp.568–572, 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] 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.
[6] 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.