文法誤り訂正モデルのエラー分析に基づく疑似データ生成の効果検証

土肥 康輔　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{doi.kosuke.de8, sudoh, s-nakamura}@is.naist.jp

1　はじめに

文法誤り訂正（Grammatical Error Correction; GEC）は，テキスト中の文法的誤りを自動的に訂正するタスクである．近年では，訂正を誤りが含まれる文から誤りが含まれない文への翻訳とみなし，ニューラル機械翻訳に基づくアプローチで取り組むことが主流となっている．ニューラル機械翻訳に基づくアプローチでは，モデルの訓練に大規模なパラレルデータが必要となるが，機械翻訳で利用可能なデータ量と比べて，GEC で利用可能なデータ量は少量である．そこで，GEC ではモデルの訓練に疑似データを用いることで，モデルの性能を向上させるということが行われている [1, 2, 3]．

本研究では，現状の GEC モデルのエラー分析に基づき，訂正性能が悪い誤りカテゴリに関する疑似誤りを訓練データに追加することで，モデルの訂正性能が向上するかを検証する．具体的には，Omelianchuk ら [4] と Grundkiewiczy ら [3] のモデル[1)]のエラー分析から訂正性能が悪いと判明した接続詞誤りに着目し，訓練に用いる学習者データにおける誤りパターンを考慮した疑似誤りを生成する．GEC では，大規模な単言語コーパスに疑似誤りを生成したデータで事前学習したモデルを，少量の学習者データで fine-tune することが一般的であるが，本研究では fine-tune に用いる学習者データに疑似誤りを生成する．

実験では，疑似誤りを追加する割合を適切に設定すれば，疑似誤りを含む学習者データで fine-tune を行った後，さらに疑似誤りを含まない学習者データで fine-tune することで，接続詞誤りの $F_{0.5}$ スコアが上昇することを確認した．また，1 回目の fine-tune のデータには疑似誤りを追加せず，2 回目のデータに疑似誤りを追加する場合でも，同様に生成する疑似誤りの割合を適切にコントロールすれば，$F_{0.5}$ スコアが上昇する可能性があることが示された．

1）現時点での最高性能のモデルと，BEA-2019 Shared Task on Grammatical Error Correction の Restricted Track での優勝システムである．

2　関連研究

疑似データの生成方法としては，逆翻訳ベースの手法や，誤りが含まれない文に疑似誤りを直接生成する手法が提案されている．

逆翻訳を用いた疑似データ生成は，ニューラル機械翻訳の文脈で Sennrich ら [5] によって提案された．[5] は，原言語と目的言語の入出力を入れ替えたモデルに目的言語の単言語コーパスを入力して疑似対訳コーパスを得る．これを GEC に応用すると，誤りが含まれない文から誤りが含まれる文を生成するようなモデルを訓練し，そのモデルに誤りが含まれない文を入力することで，誤りが含まれる文を得ることができる．Xie ら [6] はこの逆翻訳を拡張し，モデルのデコード時にノイズを加えることによって，より多くの誤りを含む文を生成できるようにした．佐藤ら [7] は，学習済みの逆翻訳モデルを特定の母語の学習者によって書かれた文で fine-tune することで，母語の影響を考慮した疑似誤りを生成する手法を提案している．

疑似誤りを直接生成する手法は，Zhao ら [8] によって提案された．この手法では，誤りが含まれない文に「置換・挿入・削除・入れ替え」の操作を行うことで誤りが含まれる文を生成する．しかし，[8] の手法では，人が犯さないような誤りを生成してしまう可能性があることが指摘されている．そこで，学習者の誤り傾向を考慮して疑似誤りを生成するということが行われている [2, 9]．

また，fine-tune によって特定のドメインに頑健な GEC モデルを作成する研究も存在する．Nădejde ら [10] は，書き手の母語情報と習熟度情報が付与された学習者コーパスを用い，一般的な GEC モデル

を 12 の母語，5 つの習熟度レベルに適応させる研究を行った．[10] では，非公開の Cambridge Learner Corpus が用いられたが，佐藤ら [7] は一般公開されている母語情報付きの学習者コーパスでモデルを fine-tune することで，3 つの母語に適応させる試みを行っている．

従来研究では全誤りカテゴリの疑似誤りを生成したデータを事前学習で用いていたのに対して，本研究では，特定の誤りカテゴリに関する疑似誤りを加えたデータでモデルを fine-tune することで，その誤りカテゴリの訂正性能を向上させることを目指す．

3 疑似データ生成手法

疑似データの生成には，特定の誤りカテゴリに関する誤りを選択的に生成するために，疑似誤りを直接生成する手法を用いる．対象の誤りカテゴリは，[3][4] のモデルのエラー分析で訂正性能が低かった接続詞である．対象となる接続詞は，{and, but, or, so}の 4 種類とし，学習者が犯す接続詞の誤りパターンを考慮するために，ERRANT[11] によって分類される「不足（Missing）・置換（Replacement）・余剰（Unnecessary）」の 3 種類の誤りタイプごとに疑似誤りの生成方法を設定する．本研究では，学習者データに疑似誤りを生成するため，疑似誤りの生成元の文にすでに誤りが含まれている場合がある．もともと存在する誤りを改変することによる影響を最小限に抑えるために，接続詞誤りがもともと含まれている文は，疑似誤りの生成対象から除外する．

接続詞が用いられている文は，確率 P で疑似誤りの生成対象とする．その文に含まれている接続詞の削除または別の接続詞への置換の操作を行うことで，それぞれ不足と置換の疑似誤りを生成する．訓練データの誤り分布分析に基づき[2)]，70%の確率で不足誤り，30%の確率で置換誤りを生成し，置換後の接続詞を選択するパラメータ $p_{repl|orig}$ は以下のように設定することとした．

$$(p_{but|and}, p_{or|and}, p_{so|and}) = (0.30, 0.60, 0.10)$$
$$(p_{and|but}, p_{or|but}, p_{so|but}) = (0.94, 0.01, 0.05)$$
$$(p_{and|or}, p_{but|or}, p_{so|or}) = (0.99, 0.01, 0.00)$$
$$(p_{and|so}, p_{but|so}, p_{or|so}) = (0.99, 0.01, 0.00)$$

なお，1 文に複数の接続詞が含まれている場合は，ランダムに選択されたひとつの接続詞が操作の対象となる．

接続詞が用いられていない文に対しては，$0.38P$ の確率でランダムな位置に接続詞を挿入することで余剰の疑似誤りを生成する．挿入する接続詞の選択は，不足・置換誤りのときと同様に訓練データの誤り分布を反映して $(p_{and|\phi}, p_{but|\phi}, p_{or|\phi}, p_{so|\phi}) = (0.65, 0.25, 0.03, 0.07)$ とした．

4 実験

4.1 データセット

実験に用いるデータは，4.2 節で後述する [4] のモデル構築に用いられたものに合わせた．事前訓練データには，PIE コーパスに疑似エラーを生成した 900 万文 [12] を用いた．訓練には，BEA-2019 Shared Task on Grammatical Error Correction で配布された Lang-8[13]，NUCLE[14]，Write & Improve + LOCNESS（W&I+L）train[15] と，CLC-FCE Dataset[16] を用いた．Lang-8 と NUCLE は，[4] でのサイズに合わせるためにサンプリングして使用した．データの 98%を訓練データ，2%を開発データとした．

評価データには，W&I+L dev，CoNLL-2013[17]，CoNLL-2014[18]，FCE test[19] を用いた．

4.2 モデル

GEC モデルには，Omelianchuk ら [4] のモデル[3)]を用いた．[4] は，入力トークンをターゲットの訂正に変換するトークンレベルの変換を新たにデザインし，GEC を系列ラベリング問題として扱っている．モデルには，BERT 系の事前学習済みモデルのエンコーダーが用いられているが，本実験では，[4] で best single model であった XLNet を用いた．ハイパーパラメータは，[4] の設定に準じた．

訓練は，疑似データによる事前学習（Stage1）と，学習者データによる 2 回の fine-tune から成る．1 回目の fine-tune（Stage2）では学習者データのうち誤りを含む文のみが用いられ，2 回目の fine-tune（Stage3）では誤りを含む文と含まない文の両方が用いられる．各データセットのサイズと，それらが訓練のどの段階で用いられたかを表 1 に示す．

4.3 実験設定

3 節で述べた疑似データ生成手法を用い，Stage2 および Stage3 で用いるデータに疑似誤りを生成した．実験は，Stage2 のみで疑似データを用いる設

2） 詳細は付録 A に掲載する．

3） https://github.com/grammarly/gector

表 1　訓練に使用したデータセット

Dataset	Sentences	Sampled	Training Stage
PIE	9,000,000	-	Stage1
Lang-8	1,037,561	947,344	Stage2
NUCLE	57,151	56,958	Stage2
W&I+L train	34,304	-	Stage2, 3
CLC-FCE	34,490	-	Stage2

表 2　モデルの構築に用いたデータの疑似誤り生成確率

Model	synthetic errors at Stage2	synthetic errors at Stage3
Baseline	-	-
Stage2_10	10%	-
Stage2_30	30%	-
Stage2_50	50%	-
Stage3_05	-	5%
Stage3_10	-	10%
Stage3_30	-	30%
Stage3_50	-	50%

定と，Stage3 のみで疑似データを用いる設定の 2 種類を行った．Stage2 での疑似誤り生成確率は，$P = (0.5, 0.3, 0.1)$，Stage3 では $P = (0.5, 0.3, 0.1, 0.05)$ とした．ベースラインには，疑似誤りを含まないデータで 2 回の fine-tune を行ったモデルを用意した．構築したモデルの一覧とその構築のために用いたデータの関係を表 2 に示す．モデルの性能は，ERRANT により算出される $F_{0.5}$ スコアにより評価した．W&I+L dev，CoNLL-2013，CoNLL-2014，FCE test のそれぞれに対し，訂正全体と接続詞訂正の $F_{0.5}$ スコアを算出した．

4.4　実験結果

実験結果を表 3 に示す．Stage3 後の接続詞訂正の $F_{0.5}$ スコアを比較すると，疑似データを導入することで，接続詞訂正の性能が向上する場合と悪化する場合の両方があることがわかる．また，CoNLL-2013 における接続詞訂正で最も高いスコア（28.41）を達成した Stage2_50 が，W&I+L dev ではベースラインよりスコアが悪化（26.79→19.23）しているように，接続詞訂正の性能がすべての評価セットでベースラインより向上しているモデルは見られなかった．これらの結果から，訂正対象のデータに応じて疑似誤り生成確率を適切に設定すれば，fine-tune に用いるデータに疑似誤りを導入する手法がモデルの訂正性能向上に効果的であることが示唆される．Stage3 で疑似誤りを導入する場合は，比較的小さい疑似誤り生成確率（W&I+L と CoNLL-2013：$P = 0.05$，CoNLL-2014 と FCE test：$P = 0.1$）を設定したときに接続詞訂正の $F_{0.5}$ スコアがベースラインから上昇しているのに対し，Stage2 で疑似誤りを導入する場合は，$P = 0.3$ または $P = 0.5$ のように大きめの生成確率を設定したときにスコアの上昇幅が大きくなっている．

4.4.1　疑似データ導入タイミング

各評価セットにおいて，接続詞訂正の $F_{0.5}$ スコアが Stage3 後に最も高くなっているのは，W&I+L dev を除いて，Stage2 で疑似誤りを導入したモデルであった．CoNLL-2013 と CoNLL-2014 においては，Stage2_50 のスコア（28.41，54.69）が最も高く，FCE test では Stage2_30 のスコア（21.28）が最も高い．Stage2_30 は，W&I+L dev の接続詞訂正においても，Stage3_05 に次ぐ 2 番目に高いスコアを達成している．Stage3 で疑似誤りを導入したモデルにも，ベースラインと比較して訂正性能が向上したものが存在するが，スコアの上昇幅は Stage2 で疑似誤りを導入したモデルよりも小さくなっている．各評価セットにおけるスコアの上昇幅の最大値を，疑似誤りの導入タイミングで比較した結果を表 4 に示す．これらの結果から，疑似誤りを含む学習者データで fine-tune を行った後，さらに疑似誤りを含まない学習者データで fine-tune するほうが，対象の誤りカテゴリに対する訂正性能が大きく改善する可能性があることが示唆される[4)]．

4.4.2　疑似データ導入の効果と影響

表 3 より，Stage2 で疑似データを導入すると，Stage2 終了時点の接続詞訂正の $F_{0.5}$ スコアはベースラインより低下することがわかる．このスコア変化は，Precision が低下する一方で，Recall が上昇することによって引き起こされている．疑似誤り導入後の接続詞訂正の Precision と Recall の値を表 5 に示す．疑似データを含まない学習者データで fine-tune したときは（S2_base），Precision が Recall より高くなっているのに対して，疑似データを導入すると Recall のほうが Precision より高くなっている．Stage3 に疑似誤りを導入した場合でも，疑似誤りの生成確率が高くなるにつれて Precision が低下する一方で Recall が上昇する傾向がみられる．fine-tune に用いるデータへの疑似誤り導入は Recall 向上に効果があり，2 回の fine-tune のうち 1 回目の段階で Recall を高めておくことが，最終的な接続詞訂正性能の向上に寄与

4)　訂正全体の $F_{0.5}$ スコアへの影響を考慮していないことに注意が必要である．疑似データ導入による影響は 4.4.2 節で述べる．

表 3 $F_{0.5}$ スコアによる訂正性能の比較．Overall は訂正全体，CONJ は接続詞訂正を表す．

	W&I+L dev		CoNLL-2013		CoNLL-2014		FCE test	
Model	Overall	CONJ	Overall	CONJ	Overall	CONJ	Overall	CONJ
After Stage2								
Baseline	46.84	28.06	43.19	19.74	55.87	28.57	52.20	21.93
Stage2_10	45.90	13.36	42.32	10.42	55.57	21.21	52.11	14.15
Stage2_30	45.50	9.08	42.45	11.36	53.74	17.72	51.61	13.83
Stage2_50	43.35	7.86	40.91	9.09	53.68	14.51	51.20	9.84
After Stage3								
Baseline	50.73	26.79	43.17	20.83	56.59	35.71	**53.35**	0.00
Stage2_10	51.02	25.57	43.18	20.83	**57.02**	38.46	53.15	11.63
Stage2_30	**51.65**	29.26	43.59	19.23	56.16	40.00	52.96	**21.28**
Stage2_50	50.76	19.23	**43.70**	**28.41**	56.03	**54.69**	**53.35**	17.44
Stage3_05	50.86	**30.49**	43.27	27.17	56.55	31.25	52.92	15.31
Stage3_10	50.89	25.74	43.09	20.16	56.47	44.64	52.81	16.23
Stage3_30	50.19	17.01	43.19	12.93	55.94	23.39	52.83	14.00
Stage3_50	49.47	9.62	42.24	7.89	55.53	16.98	51.82	10.56

表 4 疑似誤り導入タイミングによるスコア上昇幅（最大値）の比較

	Stage2 で導入	Stage3 で導入
W&I+L dev	2.47	3.70
CoNLL-2013	7.58	6.34
CoNLL-2014	18.98	8.93
FCE test	21.28	16.23

表 5 接続詞訂正の Precision と Recall の値．Stage2_XX は Stage2 終了後，Stage3_XX は Stage3 終了後の値である．

	W&I+L dev		CoNLL-2013		CoNLL-2014		FCE test	
	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.
S2_base	29.0	25.0	20.0	18.8	40.0	13.3	23.8	16.7
Stage2_10	11.5	38.6	8.9	31.3	18.9	41.2	12.5	30.0
Stage2_30	7.6	47.7	9.5	56.3	14.9	73.7	11.8	46.7
Stage2_50	6.5	52.3	7.6	50.0	12.2	65.0	8.3	40.0
S3_base	35.3	13.6	25.0	12.5	66.7	12.5	0.0	0.0
Stage3_05	33.3	22.7	26.3	31.3	37.5	18.8	17.7	10.0
Stage3_10	24.6	31.8	18.5	31.3	50.0	31.3	16.1	16.7
Stage3_30	14.7	45.5	11.1	37.5	21.1	42.1	12.7	23.3
Stage3_50	8.1	43.2	6.6	37.5	14.5	55.0	9.1	30.0

している可能性が考えられる．

また，CoNLL-2014 における Stage2_50 や FCE test における Stage2_30 のように，接続詞訂正の性能がベースラインより向上しているにも関わらず，訂正全体の性能が悪化している場合も存在する．これは，ある誤りカテゴリの疑似誤りを生成することが，別の誤りカテゴリに関するモデルの学習に影響を与えていることを示唆している．本実験では，接続詞の余剰誤りを生成する手法が問題となった可能性がある．余剰誤りを生成する対象の文は，接続詞誤りが含まれていないことは保証されているが，それ以外の誤りの有無は確認していなかった．その文のランダムな位置に接続詞が挿入されることで，もともと存在していた誤りが変質してしまったことが考えられる．

5 おわりに

本研究では，現状の GEC モデルで訂正性能が低い接続詞誤りに着目し，fine-tune に用いる学習者データに疑似誤りを追加することで，モデルの接続詞誤りの訂正性能が向上するかを検証した．実験の結果，生成する疑似誤りの割合を適切に設定すれば，fine-tune に用いる学習者データに疑似誤りを追加することで接続詞誤りの $F_{0.5}$ スコアが向上することがわかった．

本研究では接続詞誤りについてのみ実験を行ったので，別の誤りカテゴリにおいても本手法が効果的であるかを今後検証する必要がある．また，学習者データにもともと存在する誤りに影響を与えないような疑似データ生成手法の検討も必要である．例えば，誤りが存在する区間には疑似誤りを生成しないように制御する方法や，gold reference の訂正を適用後の文法的に正しい文に対して疑似誤りを生成する方法などが考えられる．疑似誤りを fine-tune データに導入することでモデルの訂正性能が向上できることが確認されれば，特定のドメインや学習者集団の特徴への適応が容易になることが期待される．

謝辞

本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

参考文献

[1] Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1236–1242, 2019.

[2] Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 213–227, 2019.

[3] Roman Grundkiewiczy, Marcin Junczys-Dowmuntz, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 252–263, 2019.

[4] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 163–170, 2020.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, 2016.

[6] Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Y. Ng, and Dan Jurafsky. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 619–628, 2018.

[7] 佐藤義貴, 和田崇史, 渡辺太郎, 松本裕治. 英語学習者の母語を考慮した文法誤り訂正のための擬似データ生成. 研究報告自然言語処理 (NL), Vol. 2020-NL-246, No. 5, pp. 1–5, 2020.

[8] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 156–165, 2019.

[9] Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 27–32, 2020.

[10] Maria Nădejde and Joel Tetreault. Personalizing grammatical error correction: Adaptation to proficiency level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 27–33, 2019.

[11] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 793–805, 2017.

[12] Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. Parallel iterative edit models for local sequence transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4260–4270, 2019.

[13] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 147–155, 2011.

[14] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31, 2013.

[15] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–75, 2019.

[16] Diane Nicholls. The Cambridge learner corpus - error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, Vol. 16, pp. 572–581, 2003.

[17] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–12, 2013.

[18] 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 *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14, 2014.

[19] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 180–189, 2011.

A　訓練データの接続詞誤り分布

訓練データ中で接続詞が用いられている文の数（470,068）に表 6 の余剰誤りの割合をかけた値と，訓練データ中で接続詞が用いられていない文の数（723,983）に不足・置換誤りの割合の和をかけた値の比から，接続詞が用いられていない文に対する疑似誤り生成確率 0.38 を得た．また，不足誤りと置換誤りの比 7:3 も表 6 から得た．余剰誤り生成のために挿入する接続詞の割合と，置換誤り生成のための置換後の接続詞の割合は，それぞれ表 7，表 8 から得た．

表 6　余剰・不足・置換誤りの分布

	誤り数	割合
余剰	5659	0.37
不足	6651	0.45
置換	2744	0.18

表 7　余剰誤りの分布

	誤り数	割合
and	3725	0.65
but	1448	0.25
or	131	0.03
so	278	0.07

表 8　置換誤りの分布

	repl				置換後の接続詞の割合
orig	and	but	or	so	(and, but, or, so)
and	-	416	874	85	(- , 0.30, 0.60, 0.10)
but	274	-	3	14	(0.94, - , 0.01, 0.05)
or	647	4	-	0	(0.99, 0.01, - , 0.00)
so	51	24	0	-	(0.99, 0.01, 0.00, -)

B　事例分析

疑似データを用いたことによるモデル出力の変化を表 9 に示す．不足誤りと置換誤りについては，and と or が関わる誤りにおいて疑似誤りを用いたモデルで改善が見られた．しかし，余剰誤りについては，あまり改善が見られなかった．改善が見られた誤りは，疑似データ生成の確率パラメータが大きく設定されているものである．多数の疑似誤りが生成された誤りの訂正性能が高まる傾向は，[2] で報告されている結果と一致する．また，ベースラインでは正しく訂正できていたが疑似データを用いたモデルでは訂正に失敗した例，不必要な訂正をしてしまった例を表 10 に示す．

表 9　疑似データを用いたことによるモデル出力の変化例

Dataset	Model	Sentence
W&I+L dev	original	It was a dark night it was raining until a big ...
	gold	It was a dark night **and** it was raining when a big ...
	baseline	It was a dark night **.** **It** was raining when a big ...
	Stage3_05	It was a dark night **and** it was raining until a big ...
CoNLL-2014	original	They may set a bias on this person even abandon his or her .
	gold	They may discriminate against this person **or** even abandon him or her .
	baseline	They may set a bias on this person , even abandon him or her .
	Stage2_30	They may set a bias on this person **or** even abandon him or her .
	Stage3_05	They may set a bias on this person **and** even abandon him or her .
FCE test	original	If you have more questions about the conference **and** something else , ...
	gold	If you have more questions about the conference **or** anything else , ...
	baseline	If you have more questions about the conference **and** anything else , ...
	Stage2_10	If you have more questions about the conference **or** anything else , ...
W&I+L dev	original	... source of energy does n't always maintain at the constant level , **but** someday it will be run out .
	gold	... source of energy does n't always remain at a constant level , **and** someday it will run out .
	baseline	... source of energy does n't always stay at a constant level , **but** someday it will run out .
	Stage2_30	... source of energy does n't always stay at a constant level , **but** someday it will run out .
FCE test	original	I hope you will be happy with our conference and party **and** etc .
	gold	I hope you will be happy with our conference and party etc .
	baseline	I hope you will be happy with our conference , party **,** etc .
	Stage2_30	I hope you will be happy with our conference and party **,** etc .

表 10　疑似データを用いたモデルで訂正結果が悪くなった例

Dataset	Model	Sentence
W&I+L dev	original	Although , one day later the headmaster found out the truth through CCTV , **but** we refused ...
	gold	Although , one day later the headmaster found out the truth through CCTV , we refused ...
	baseline	Although , one day later , the headmaster found out the truth through CCTV , we refused ...
	Stage2_30	Although , one day later , the headmaster found out the truth through CCTV , **but** we refused ...
W&I+L dev	original	Chinese American Literature is philisophical / literal because ...
	gold	Chinese American Literature is philisophical / literal because ...
	baseline	Chinese American Literature is philisophical / literal because ...
	Stage2_50	Chinese **and** American Literature is philisophical / literal because ...