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

NAIST

文法誤り訂正（GEC）のタスク

- テキスト中の文法誤りを自動的に訂正

入力：文法誤りが含まれる文

Travel by bus is exspensive , bored and annoying .

出力：文法的に正しい文

Travelling by bus is expensive , boring and annoying .

GECへのアプローチ

- ニューラル機械翻訳に基づくアプローチが主流
  - 原言語：文法誤りが含まれる文
  - 目的言語：文法的に正しい文

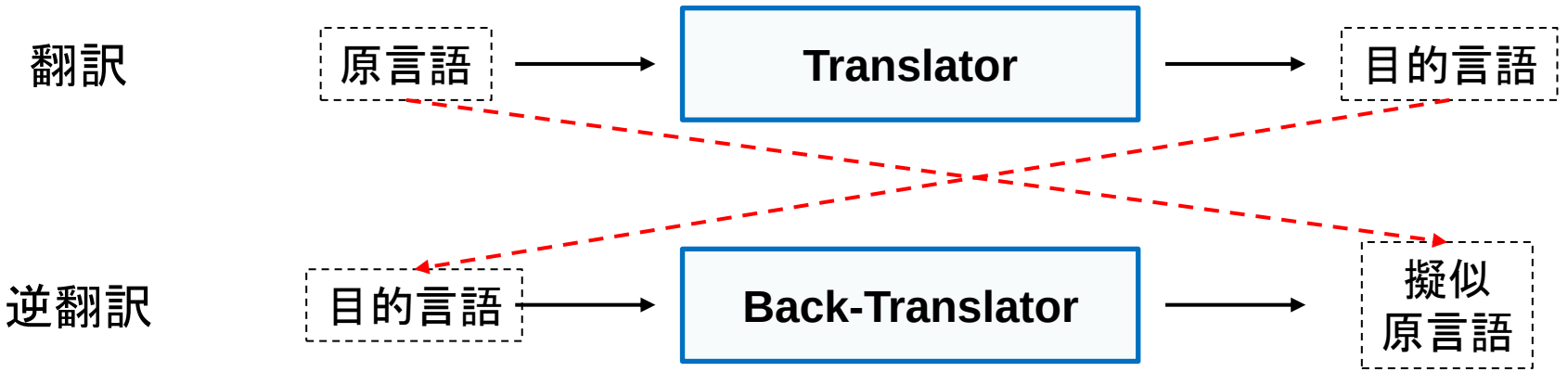
- 問題点
  - GECで利用可能なデータ量が少ない

→ 疑似データを利用

[Kiyono+ 2019, Choe+ 2019, Grundkiewiczy+ 2019]

関連研究：疑似データ生成手法（1）

- 逆翻訳 [Sennrich+ 2016]



GECでは…
文法的に正しい文から，文法誤りを含む文を生成する

- ノイズを用いた逆翻訳 [Xie+ 2018]
  - ビームサーチ時の仮説のスコアにノイズを加える

→ より多くの誤りを含む文が生成可能

関連研究：疑似データ生成手法（2）

- 疑似誤りを直接生成する手法 [Zhao+ 2019]
  - 文法的に正しい文に「置換・挿入・削除・入れ替え」の操作を行う

(correct)	I am looking forward to receiving your answer!
(realistic)	I am looking forward for receiving your answer!
(unrealistic)	I am looking forward mountain receiving your answer!

- 学習者の誤り傾向を考慮した手法 [Choe+ 2019, Takahashi+ 2020]
  - 疑似誤りを直接生成する手法では，人が犯さないような誤りを生成する可能性がある

本研究での疑似データの利用法

- 既存のGECモデルのエラー分析に基づき，疑似誤りを生成

一般的な利用法：
全誤りカテゴリの疑似誤りを
事前学習データに生成

事前学習
単言語コーパスに疑似誤りを生成したデータ

↓

Fine-tune
学習者データ

本研究：
特定の誤りカテゴリの疑似誤りを
fine-tuneデータに生成

本研究の概要

- 特定の誤りカテゴリの疑似誤りをfine-tuneデータに生成
  → 疑似誤りを直接生成する手法
  → GECモデルの性能が向上するか？

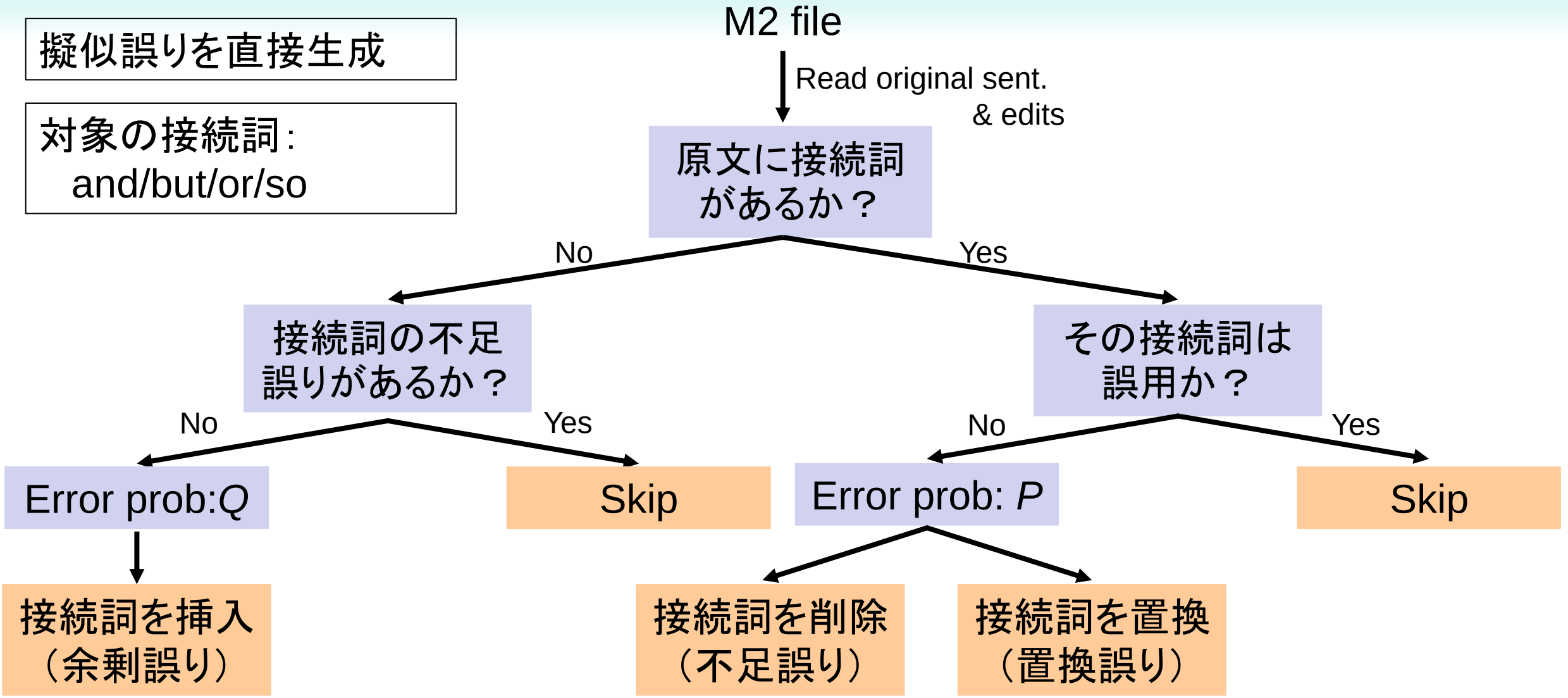
- 対象 : 接続詞誤り {and/but/or/so}
  – 既存のGECモデル*1 で訂正性能がよくない

*1 [Omelianchuk+ 2020]: 現時点でSOTAのモデル
[Grundkiewiczy+2019]: BEA-2019 (Restricted Track) の優勝システム

本研究の疑似誤り生成手法

擬似誤りを直接生成

対象の接続詞：
and/but/or/so



訓練データでの誤り傾向分析

- 余剰・不足・置換誤りの割合

	# of err	%
余剰	5659	0.37
不足	6651	0.45
置換	2744	0.18

0.37 * 接続詞を含む文の数：470,068 = 173,925.16

0.63 * 接続詞を含まない文の数：723,983 = 456,109.29

余剰 / 不足+置換= 0.38

不足：置換 = 0.7 : 0.3

- 余剰誤り

	# of err	%
and	3725	0.65
but	1448	0.25
or	131	0.03
so	278	0.07

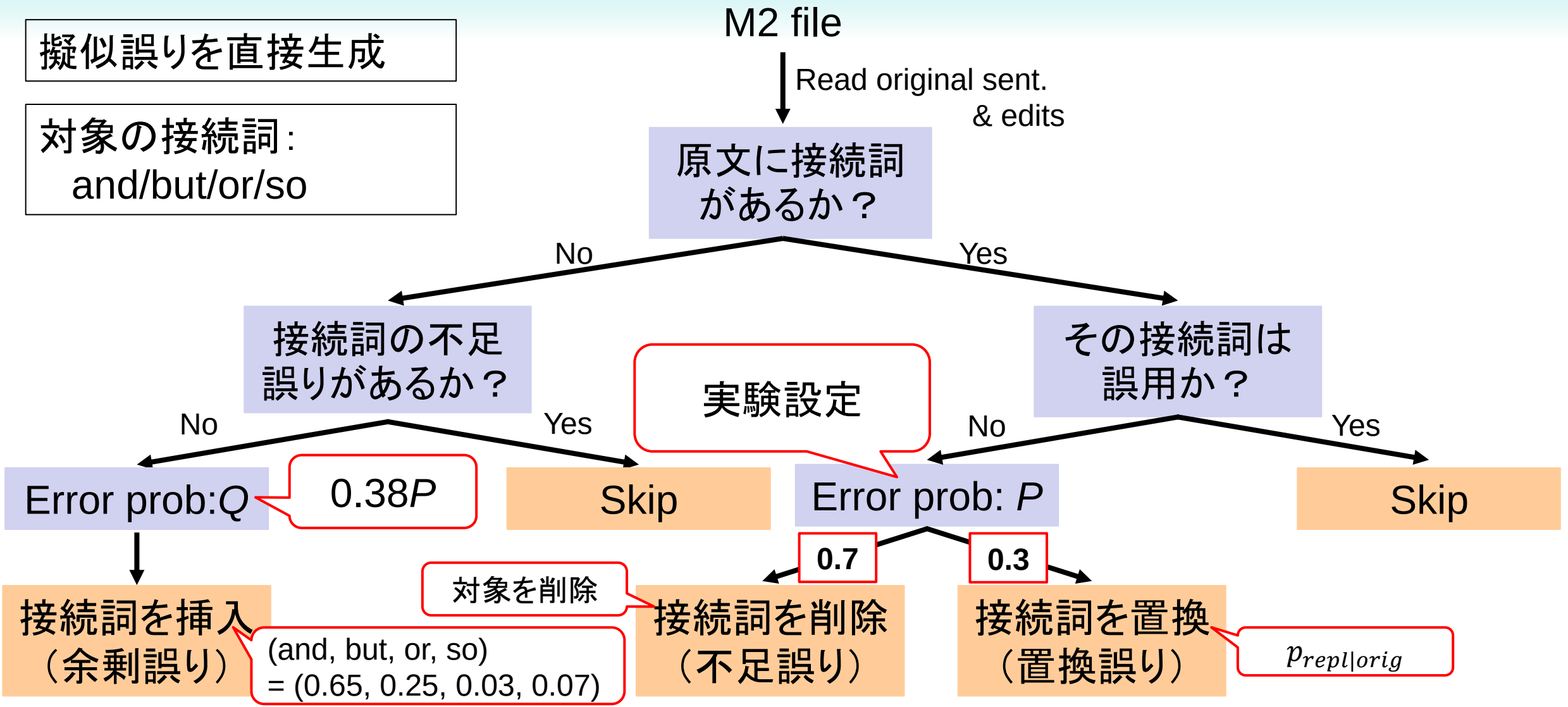
- 置換誤り

	repl				$p_{repl\|orig}$
orig	and	but	or	so	$(p_{and\|orig}, p_{but\|orig}, p_{or\|orig}, p_{so\|orig})$
and	-	416	874	85	(0.00, 0.30, 0.60, 0.10)
but	274	-	3	14	(0.94, 0.00, 0.01, 0.05)
or	647	4	-	0	(0.99, 0.01, 0.00, 0.00)
so	51	24	0	-	(0.99, 0.01, 0.00, 0.00)

本研究の疑似誤り生成手法

擬似誤りを直接生成

対象の接続詞：
and/but/or/so



実験設定：GECモデル

- モデル
  - GECToR [Omelianchuk+ 2020][*1]
  - 系列ラベリング問題としてGECにアプローチ
  - BERT系の事前学習済みモデルのエンコーダーを利用
    - XLNet (= [Omelianchuk+ 2020]でbest single model)
  - 訓練ステージ
    - Stage1: 疑似データによる事前学習
    - Stage2: 学習者データのうち，誤りを含む文のみでfine-tune
    - Stage3: 学習者データ全体でfine-tune

    *本研究では，Stage2/3にも疑似データを生成した

*1 https://github.com/grammarly/gector

データ

- 訓練データ
  - [Omelianchuk+ 2020]と同じ
  - 98% = 訓練 / 2% = 開発
- 評価データ
  - W&I+L dev, CoNLL-2013, CoNLL-2014, FCE test
  - ERRANT [Bryant+ 2017] により算出される $F_{0.5}$スコア（訂正全体・接続詞訂正）で評価

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

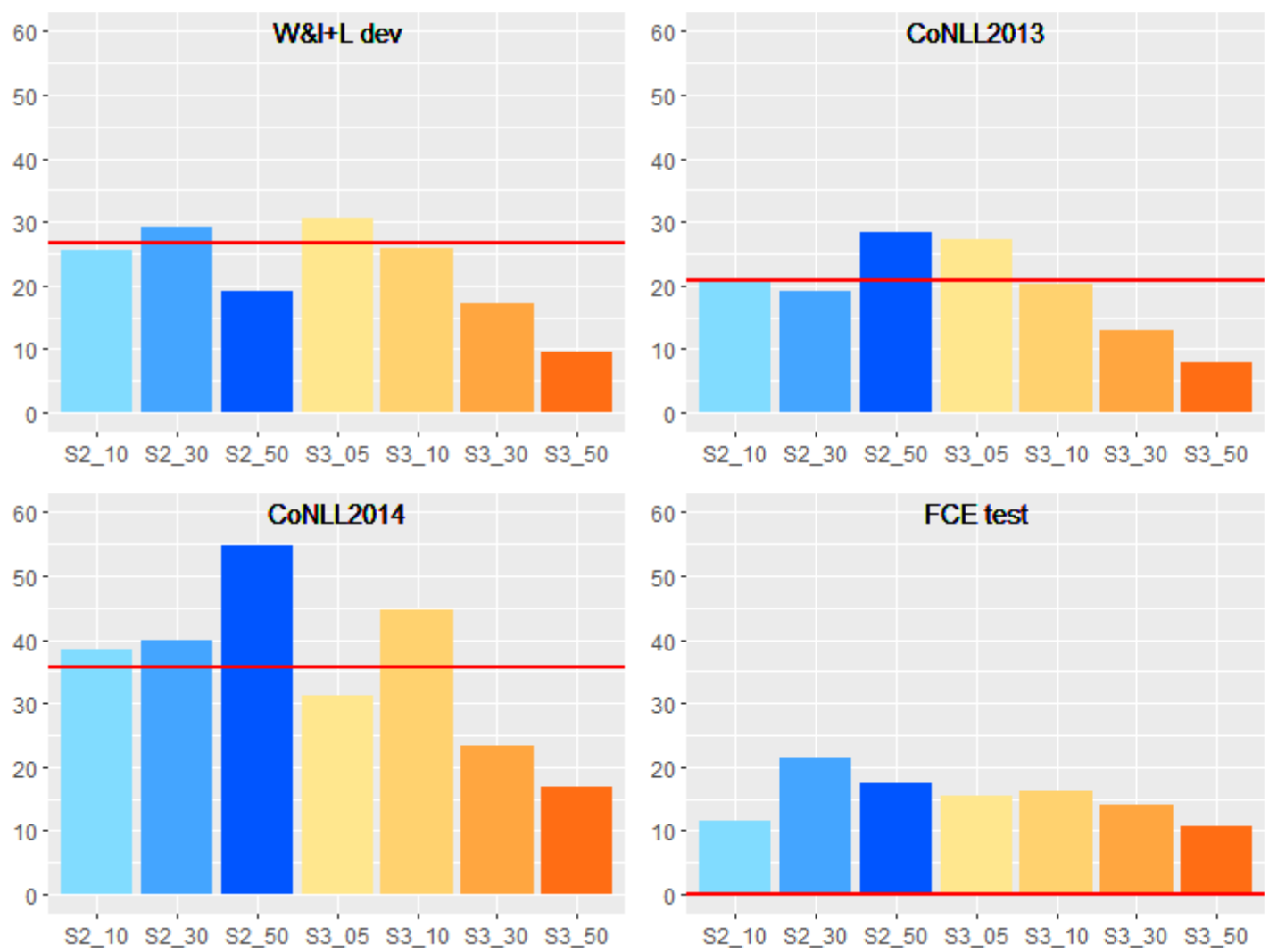
Dataset	Sentences
W&I+L dev	4,384
CoNLL-2013	1,381
CoNLL-2014	1,312
FCE test	2,695

実験設定

- Stage2 または Stage3で疑似誤りを生成
  - Stage2: P = (0.1, 0.3, 0.5)
  - Stage3: P = (0.05, 0.1, 0.3, 0.5)

	Model							
Stage	Baseline	S2_10	S2_30	S2_50	S3_05	S3_10	S3_30	S3_50
Stage1 (pre-training)	PIE corpus (900M sentences) w/ synthetic errors							
Stage2 (fine-tune)	-	10%	30%	50%	-	-	-	-
Stage3 (fine-tune)	-	-	-	-	5%	10%	30%	50%

結果：接続詞訂正の$F_{0.5}$スコア（Stage3後）



- スコアは向上/悪化の両方がある
- すべての評価データでスコアが向上しているモデルは存在しない

  →訂正対象のデータに応じて誤り生成確率を適切に設定する必要がある
- Stage2では比較的大きい*P*がよいが，Stage3では小さい*P*のほうがよい

疑似誤り導入のタイミング

- Stage2で導入するほうが効果的な可能性
  - S2_XXが最も高いスコアを達成している（W&I+L devを除く）

	W&I+L dev	CoNLL2013	CoNLL2014	FCE test
最高スコアのモデル	S3_05	S2_50	S2_50	S2_30

  - S2_XXのほうがスコアの上昇幅が大きい
    - 接続詞訂正の$F_{0.5}$スコアの上昇幅（最大値）の比較

	S2_XX	S3_XX
W&I+L dev	2.47	3.70
CoNLL2013	7.58	6.34
CoNLL2014	18.98	8.93
FCE test	21.28	16.23

疑似誤り導入の効果

- (S2_XX) Stage2後の接続詞訂正の$F_{0.5}$スコアは，ベースラインより悪化している
  - Precision = 低下 / Recall = 上昇
- (S3_XX) 擬似誤り生成確率が高いほど，Recallが高くなる傾向

W&I+L dev（Stage2後）

Model	Precision	Recall	$F_{0.5}$
Baseline	28.95	25.00	**28.06**
S2_10	11.49	38.64	13.36
S2_30	7.55	47.73	9.08
S2_50	6.48	52.27	7.86

W&I+L dev（Stage3後）

Model	Precision	Recall	$F_{0.5}$
Baseline	35.29	13.64	26.79
S3_05	33.33	22.73	**30.49**
S3_10	24.56	31.82	25.74
S3_30	14.71	45.45	17.01
S3_50	8.05	43.18	9.62

疑似誤り導入の効果

- Fine-tuneに用いるデータに疑似誤りを導入することは，Recallの上昇に効果がある
- Stage2でRecallを高めておくことが，最終的な$F_{0.5}$スコアの向上に寄与する可能性
- [問題点]接続詞訂正の $F_{0.5}$スコアが向上しているにも関わらず，訂正全体の$F_{0.5}$スコアが悪化する場合がある
  - 例：CoNLL2014におけるS2_30（接続詞：+4.29 / 全体：-0.43）
  - 「余剰誤り」生成手法が影響？

まとめ

- 既存のGECモデルの誤り分析に基づき，接続詞誤りに着目
- 適切な疑似誤り生成確率P のもとで接続詞の疑似誤りをfine-tuneデータに導入することで，GECモデルの性能が向上した

課題

- 接続詞誤りしか検証していない
- 疑似誤り生成手法が他の誤りカテゴリの学習に影響している

  →他の誤りカテゴリで検証する

  →疑似誤り生成手法をより洗練させる

参考文献

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. 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)*, pages 793-805.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A Neural Grammatical Error Correction System Built On Better Pre-training and Sequential Transfer Learning. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213-227.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252-263.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. 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)*, pages 1236-1242.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163-170.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 86-96.

参考文献

Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. 2020. 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*, pages 27-32.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Y. Ng, and Dan Jurafsky. 2018. 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)*, pages 619-628.

Wei Zhao, LiangWang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. 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)*, pages 156-165.

Appendices

結果：$F_{0.5}$スコア（Stage3後）

	W&I+L dev		CoNLL2013		CoNLL2014		FCE test	
Model	Overall	CONJ	Overall	CONJ	Overall	CONJ	Overall	CONJ
Baseline	50.73	26.79	43.17	20.83	56.59	35.71	**53.35**	0.00
S2_10	51.02	25.57	43.18	20.83	**57.02**	38.46	53.15	11.63
S2_30	**51.65**	29.26	43.59	19.23	56.16	40.00	52.96	**21.28**
S2_50	50.76	19.23	**43.70**	**28.41**	56.03	**54.69**	**53.35**	17.44
S3_05	50.86	**30.49**	43.27	27.17	56.55	31.25	52.92	15.31
S3_10	50.89	25.74	43.09	20.16	56.47	44.64	52.81	16.23
S3_30	50.19	17.01	43.19	12.93	55.94	23.39	52.83	14.00
S3_50	49.47	9.62	42.24	7.89	55.53	16.98	51.82	10.56

結果：$F_{0.5}$スコアの増減（Stage3後）

	W&I+L dev		CoNLL2013		CoNLL2014		FCE test	
Model	Overall	CONJ	Overall	CONJ	Overall	CONJ	Overall	CONJ
Baseline	-	-	-	-	-	-	-	-
S2_10	0.29	-1.22	0.01	0.00	0.43	2.75	-0.2	11.63
S2_30	0.92	2.47	0.42	-1.60	-0.43	4.29	-0.39	21.28
S2_50	0.03	-7.56	0.53	7.58	-0.56	18.98	0.00	17.44
S3_05	0.13	3.70	0.10	6.34	-0.04	-4.46	-0.43	15.31
S3_10	0.16	-1.05	-0.08	-0.67	-0.12	8.93	-0.54	16.23
S3_30	-0.54	-9.78	0.02	-7.90	-0.65	-12.32	-0.52	14.00
S3_50	-1.26	-17.17	-0.93	-12.94	-1.06	-18.73	-1.53	10.56

結果：$F_{0.5}$スコア（Stage2後）

	W&I+L dev		CoNLL2013		CoNLL2014		FCE test	
Model	Overall	CONJ	Overall	CONJ	Overall	CONJ	Overall	CONJ
Baseline	46.84	28.06	43.19	19.74	55.87	28.57	52.2	21.93
S2_10	45.90	13.36	42.32	10.42	55.57	21.21	52.11	14.15
S2_30	45.50	9.08	42.45	11.36	53.74	17.72	51.61	13.83
S2_50	43.35	7.86	40.91	9.09	53.68	14.51	51.20	9.84

接続詞訂正のPrecision/Recall

Model	W&I+L dev		CoNLL2013		CoNLL2014		FCE test	
	P	R	P	R	P	R	P	R
S2_base	28.95	25.00	20.00	18.75	40.00	13.33	23.81	16.67
S2_10	11.49	38.64	8.93	31.25	18.92	41.18	12.50	30.00
S2_30	7.55	47.73	9.47	56.25	14.89	73.68	11.76	46.67
S2_50	6.48	52.27	7.55	50.00	12.15	65.00	8.28	40.00
S3_base	35.29	13.64	25.00	12.50	66.67	12.50	0.00	0.00
S3_05	33.33	22.73	26.32	31.25	37.50	18.75	17.65	10.00
S3_10	24.56	31.82	18.52	31.25	50.00	31.25	16.13	16.67
S3_30	14.71	45.45	11.11	37.50	21.05	42.11	12.73	23.33
S3_50	8.05	43.18	6.59	37.50	14.47	55.00	9.09	30.00

S2_XX: Stage2の値
S3_XX: Stage3後の値

Error analysis of existing GEC models

- Error categories whose $F_{0.5}$ score is low
  - Errors that have variability for correction
    - Adverb (ADV), adjective (ADJ), noun (NOUN), verb (VERB) word order (WO)
  - Errors related to meaning
    - Adverb (ADV), adjective (ADJ), conjunction (CONJ), noun (NOUN), verb (VERB)

[Grundkiewicz+ 2019]		[Omelianchuk+ 2020]	
Err category	$F_{0.5}$ (avg.)	Err category	$F_{0.5}$ (avg.)
CONJ	26.81	OTHER	21.12
OTHER	29.62	CONJ	22.03
ADV	32.02	ADV	25.53
NOUN	35.98	CONTR	26.07
CONTR	37.01	NOUN	28.03
VERB	39.39	ADJ	32.64
PRON	41.06	VERB	33.06
WO	45.73	WO	35.37
PUNCT	46.10	PART	39.43
VERB:TENSE	46.52	PRON	45.09

Output examples of the models (1/3)

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

Output examples of the models (2/3)

Dataset	Model	Sentence
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 .

Output examples of the models (3/3)

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 ...