破滅的忘却を防ぐ最適化関数を用いた構文情報の事前学習

岩本 蘭 [1,2] 吉田 一星 [1] 金山 博 [1] 大湖 卓也 [1]
[1] 日本アイ・ビー・エム株式会社 東京基礎研究所 [2] 慶應義塾大学
ran.iwamoto1@ibm.com {issei,hkana,ohkot}@jp.ibm.com

概要

構文構造は複雑な文や長い文の中の重要な情報に目を向ける手助けとなる。しかし BERT などの事前学習済モデルは明示的な構文制約を与えておらず、感情分析タスクなどで構文的に不自然な出力が見られる。その問題への対処として構文知識を事前学習させる研究が行われている。本論文では事前学習済 BERT がもつ意味表現を保持しつつ構文知識を加えるために、破滅的忘却を防ぐ 2 つの最適化関数 (Elastic Weight Consolidation と Gradient Surgery) を用いて 4 つの構文タスクをそれぞれ追加学習させた。追加学習済モデルを用いて GLUE の 3 つのタスク (CoLA, RTE, MRPC) を解き、意味と構文両方の情報を持つモデルが高性能を発揮することを示した。

1 はじめに

事前学習済言語モデル（以降、モデルと呼ぶ）は近年の言語処理タスクの性能向上に大きく寄与している。それらのモデルは品詞や依存構造などの構文情報を大まかに捉えているが [3]、実際には応用タスクを解くための構文情報がまだ不足している [20]。

構文構造を明示的/暗黙的に組み込んだ言語処理モデルは機械読解 [24] や言語理解 [23]、翻訳 [2] などで高い性能を発揮している。それらの研究ではモデルに構文構造を読み込むモジュールを明示的に組み込む。しかし既存の応用タスクの多くは Hugging Face [19] などが配布している形式のモデルの読み込みを前提としており、他のタスクへの活用が難しいという側面がある。そこでモデルの再利用性の向上を目的とし、構文構造の追加学習のみを行うモデルが台頭している [16, 21]。本論文では近年の流れに倣い、BERT [4] への構文構造の追加事前学習を行い、後段タスクで使いやすいモデルを作成した。学習済モデルに別のタスクを学習させると以前学んだことを忘れる破滅的忘却 [5] という問題が起こりやすいため、破滅的忘却を防ぐ最適化関数を用いた。

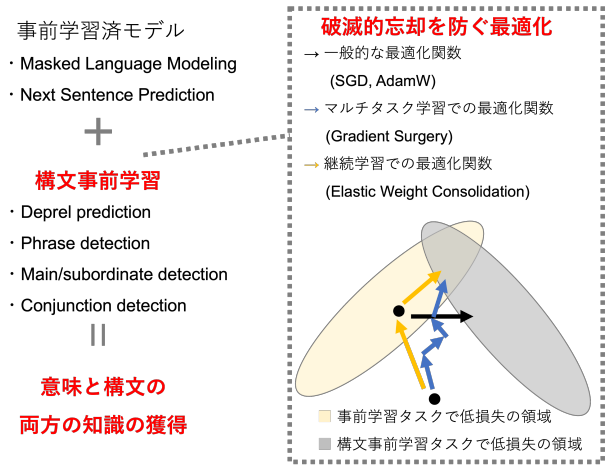

図 1 構文事前学習の概要。破滅的忘却を防ぐ最適化関数を用いて構文情報を追加学習し、モデルが元々持つ意味表現を損なわず構文情報を獲得する。

構文事前学習の概要を図 1 に示す。本論文の主な貢献は以下の 3 つである。

- 事前学習済 BERT に 4 つの異なる構文情報を新たに組み込み (構文学習)、後段タスクでの性能を比較した。
- 構文学習の際に、既存の知識を忘れてしまう破滅的忘却を防ぐ最適化関数を用いて、意味情報と構文知識の両方を持つモデルを作成した。
- GLUE の 3 タスクでの実験で、事前学習済 BERT が元々持つ意味情報と、追加学習で得た構文知識をバランスよく含むモデルが高い性能を発揮することを示した。

2 構文構造の事前学習

本節では学習させる構文情報の種類について述べる。既存研究では係り受けの予測 [18] や、依存関係の距離、兄弟関係の予測 [16] など主に 2 つの語の関係を用いている。実応用、例えば名詞句の抽出 [6] や感情分析 [7] などでは主節/従属節の関係や並列構造を従来モデルを用いてうまく扱えていない。タスクに応じて必要な構文情報をモデルに反映させるに

テキスト			ラベル1	ラベル2			
id	form	PoS	head	deprel	phrase	main/sub	conj
1	We	PRON	3	nsubj	nsubj	main	other
2	still	ADV	3	advmod	advmod	main	other
3	have	VERB	0	root	root	main	other
4	the	DET	5	det	obj	main	child
5	traders	NOUN	3	obj	obj	main	conj
6	and	CCONJ	7	cc	obj	main	cc
7	books	NOUN	5	conj	obj	main	conj
8	that	PRON	10	obj	obj	subordinate	other
9	you	PRON	10	nsubj	nsubj	subordinate	other
10	provided	VERB	5	acl	acl	subordinate	other
11	last	ADJ	12	amod	obl	subordinate	other
12	week	NOUN	10	obl	obl	subordinate	other
13	.	PUNCT	3	punct	punct	main	other

図 2　構文事前学習タスクの例。Tian らのシステム [16] を用い、4 つの事前学習ではラベル 1 と 2 を同時に学習する。青い部分はフレーズや主節、並列構造を持つ。

は、多様性に富んだ構文学習を行う必要がある。

そこで我々は構文事前学習として、1) deprel prediction 2) phrase detection, 3) main/subordinate detection, 4) conjunction detection を行い、どのような構文情報が応用タスクに有用かを調べた。

2.1　構文事前学習システム

構文学習のために Tian らのシステム [16] を用いた。システムは係り受けを予測する dependency masking (DM) と関係名を予測する masked dependency prediction (MDP) を同時学習する構造を持つ。本論文では MDP 側を我々のタスクに変更する。

2.2　構文事前学習タスク

本節では図 2 の 4 つの事前学習について述べる。それぞれ文中の単語の係り受けを予測する部分は同じとし (図 2 のラベル 1 に相当)、図 2 のラベル 2 の部分が各タスクによって異なる。

Deprel prediction (deprel)　依存構造解析と同様であり、構文事前学習のベースラインタスクとする。

Phrase detection (phrase)　Basirat ら [1] から着想を得たタスクである。彼らはフレーズのような単位である nucleus を用いて依存構造解析の性能を向上させた。本論文でもそれに倣いフレーズ間の関係 (フレーズの主辞の deprel) を予測する。

Main/subordinate detection (main/sub)　Nikolaev ら [12] の、BERT が従属節を検出可能かどうかを調査するタスクを参考にし、主節と従属節を予測する。

Conjunction detection (conj)　A and B といった並列構造を予測する。具体的には A, B それぞれの主辞 (conj) とその中の子要素 (child)、and 等の並列接続詞 (cc)、それ以外 (other) のラベルを予測する。

3　破滅的忘却を防ぐ最適化

本節ではモデルの意味情報を保持しつつ構文知識を追加するための 2 つの最適化関数を紹介する。

モデルに対してタスクを連続で学習させると、最初の方のタスクの性能が著しく低下する破滅的忘却 [5] という問題が生じる。事前学習済モデルでも破滅的忘却を防ぐ様々な研究が提案されており [8]、我々はマルチタスク学習、特に強化学習や翻訳の分野でよく用いられる最適化手法 Gradient Surgery (GS)[22] と、継続学習で一般的な最適化である Elastic Weight Consolidation (EWC)[9] を用いた。

3.1　Gradient Surgery (GS)

Gradient Surgery は複数タスクを同時に学習するマルチタスク学習で、勾配の対立を解消する最適化手法である。例えばある 2 つのタスクで勾配が逆方向になる場合、片方の勾配をもう一方の勾配の直行平面に射影し、もう一方の勾配と対立している部分を消す。その後 2 つの勾配を足し合わせる。

3.2　Elastic Weight Consolidation (EWC)

EWC はタスクを順に学習する継続学習で最初のタスクの最適解を探した後 (または学習済モデルを用いる時) にそのタスクと次のタスク両方で高性能を出すパラメータを探す最適化手法である。Fisher 情報行列を用いて、最初のタスクで重要なパラメータはなるべく更新せず、重要でないパラメータの更新重みを大きくする。GS を用いるマルチタスク学習では最初のタスクのデータを大量に必要とするのに対し、継続学習で EWC を用いる際には少量のデータのみでパラメータの重要度を計算できる。

4　実験 1: 構文事前学習

本実験では事前学習済モデルが意味情報を残したまま構文情報をどの程度学習可能かを検証する。

4.1　設定

構文学習には 2 節で述べた Tian らのシステム [16] を、事前学習済モデルとして bert-base-cased[1)] を用いた。構文学習のデータは UD-EWT v2.10 [15] から作成し、train/dev/test の割合は変えていない。単語数が 5 未満の文、PoS タグ “X” や関係ラベル “dep” を含む文を除去した。実験に用いた文数を表 2 に示す。

1)　https://huggingface.co/bert-base-cased

表 1　構文構造の事前学習での、構文タスクと最適化関数の違いによる構文/MLM タスクの性能比較。

additional	optimization	Syntax Head			Syntax Label			Avg.	MLM
pretrain	function	rec	prec	f1	rec	prec	f1	f1	PPL
no	-	-	-	-	-	-	-	-	29.10
deprel	AdamW	98.02	96.04	96.82	96.90	95.37	95.63	**96.22**	702.87
	SGD	97.86	95.86	96.64	96.90	95.37	95.63	96.13	32392.10
	GS	97.30	94.75	95.69	96.48	94.93	95.18	95.43	**10.44**
	EWC	97.43	94.70	95.56	96.03	94.40	94.71	95.13	25.63
phrase	AdamW	97.93	97.00	97.29	97.69	97.00	97.21	**97.25**	1360.61
	SGD	97.19	95.20	95.89	95.64	95.44	95.36	95.62	2587.11
	GS	97.58	95.80	96.53	95.08	95.21	95.02	95.78	**9.68**
	EWC	97.70	95.74	96.51	94.79	94.72	94.66	95.58	24.17
main/subordinate	AdamW	98.14	96.98	97.40	99.67	98.96	99.31	**98.35**	720.49
	SGD	96.90	94.80	95.49	100.00	100.00	100.00	97.74	1471.63
	GS	95.70	93.86	94.16	94.93	95.56	95.24	94.70	**11.16**
	EWC	97.57	95.22	96.07	94.52	94.52	94.52	95.29	23.59
conj	AdamW	91.30	90.44	89.90	99.70	99.22	99.45	**94.68**	2971.79
	SGD	83.81	83.65	82.57	97.49	95.79	96.53	89.55	1412.65
	GS	89.57	88.19	87.75	96.46	93.93	94.92	91.33	**9.56**
	EWC	89.57	88.19	87.75	96.13	93.15	94.26	91.00	23.78

表 2　実験に使用した UD-EWT コーパスの文数.

	train	dev	test
元コーパス	12543	2001	2077
ノイズ除去後	10271	1471	1485

最適化関数は 3 節で述べた破滅的忘却を防ぐ GS、EWC と、比較手法として AdamW [10] と SGD [13] を用いた。GS と EWC を用いた学習では構文学習の 1 ステップごとに WikiText2 [11] からランダムに 100 文を選び MLM (Masked Language Modeling) の勾配を求めた。GS ではその勾配を構文学習の勾配に足す、つまり少量の MLM データでのマルチタスク学習を行なった。EWC ではパラメータの重要度計算に勾配を用いただけで、MLM の追加学習はしていない。

bert-base-cased を構文学習なしのベースラインとし、4 つの構文学習をそれぞれ行なった。BERT の元々の意味情報を保持しているか確かめるため、構文学習後に全てのモデルに対して WikiText2 を用いて MLM を測定した。MLM の評価尺度として擬似対数尤度スコア (PPL) [14] を、構文タスクの評価尺度として適合率、再現率、F1 スコアを用いた。学習率は{1e-4,1e-5,1e-6}、epoch 数は{50,70,100}を用い、test データで Syntax Head と Syntax Label 予測の平均 F1 スコアが一番高いモデルの値を載せた。

4.2　結果

結果を表 1 に示す。GS と EWC のみ MLM 事前学習の知識を保ったまま (PPL が低いまま) 構文学習ができたのに対し、AdamW と SGD は破滅的忘却が起こり PPL が増大した。AdamW はどの構文タスクでも最も良い平均 F1 スコアを達成したが、特に deprel タスクにおいては他の最適化との F1 スコアの差はほぼない。AdamW と SGD は target タスク (ここでは構文学習) の性能向上に特化するので、直前のタスクの性能が低くなる (PPL が高くなる)。PPL と応用タスクの関係については栗林ら [25] も触れている。GS と EWC では PPL が低く、構文学習よりも前に身につけた意味的な知識を保っていることがわかる。

タスクごとの違いについても述べる。AdamW での Syntax Label のスコアの違いを比較すると、main/subordinate タスクや conj タスクは Syntax label の種類が少ないため F1 スコアが高いが、それ以外のタスクは Syntax label の種類が多く F1 スコアが低い。この結果はデータの性質から予測可能である。注目すべきは Syntax Head で、4 タスクで正解が同じなのにも関わらず、もう片方のタスクとの相性によってスコアが異なる。例えば、conj と head の組み合わせは他に比べて F1 スコアが低い。

表 3　構文事前学習タスクと最適化関数のそれぞれの違いに基づく GLUE タスクの性能比較。構文事前学習タスクごとの最も高いスコアを下線、それぞれの GLUE タスクにおいての最高スコアを太字で示す。

Syntax Pretrain		GLUE Tasks				Syntax Pretrain Info	
Task	Optimizer	CoLA (mc)	RTE (acc)	MRPC (acc/f1)	Avg.	Avg.F1	PPL
no	-	60.10	64.62	84.56 / 89.23	70.54	-	29.10
deprel	AdamW	57.60	66.43	78.19 / 85.85	68.68	95.74	605.63
	SGD	57.80	67.15	85.05 / 89.78	70.79	24.87	137.17
	GS	59.07	66.07	**86.52** / 90.53	71.22	79.96	20.05
	EWC	60.57	66.07	85.54 / 89.92	71.45	77.35	26.98
phrase	AdamW	55.98	67.15	78.68 / 86.21	68.53	93.46	320.29
	SGD	58.30	67.87	84.31 / 89.33	71.00	25.60	112.02
	GS	60.86	64.98	84.56 / 89.23	70.91	59.96	20.72
	EWC	60.32	67.15	86.28 / **90.54**	71.96	58.21	23.56
main/subordinate	AdamW	55.73	66.07	77.94 / 85.44	67.83	96.44	298.95
	SGD	57.31	67.51	85.29 / 89.90	70.81	48.98	133.98
	GS	59.10	67.51	85.29 / 89.83	71.39	72.51	20.71
	EWC	60.09	68.23	84.80 / 89.49	71.82	72.10	23.12
conj	AdamW	58.29	63.90	77.45 / 85.16	67.83	94.03	321.55
	SGD	59.31	66.07	83.82 / 88.81	70.56	45.42	160.63
	GS	59.56	**68.59**	85.54 / 90.02	71.98	78.98	13.96
	EWC	**61.10**	68.23	85.05 / 89.61	**72.22**	71.48	22.30

5　実験 2: GLUE

構文知識が言語理解に及ぼす効果を、言語理解のベンチマーク GLUE [17] の中で構文構造が有用と思われるタスク (CoLA, RTE, MRPC) で評価する。CoLA はある英文の文法が正しいかどうかを、RTE は 2 つの文の含意関係を、MRPC は 2 つの文が同じ意味かを判定する 2 値分類タスクである。

5.1　設定

訓練時の epoch 数は 5 に設定し、評価指標は GLUE に倣い、CoLA では Matthews correlation, RTE では精度、MRPC では精度と F1 スコアを用いた。また, 3 タスクの評価値のマクロ平均をとった。[2)] 次節で示す結果は表 1 のモデルとは異なり、GLUE の 3 タスクの平均スコアが一番高いモデルの結果を載せている。そのため構文学習のスコアも記している。

本論文では最適化と事前学習の違いによる性能の違いを開発データのみで評価した。鈴木ら [26] が述べたように、リーダーボードに多数の投稿をすることは評価データの傾向の判明につながるため、本実験では開発データでの性能比較で十分と判断した。

2)　MRPC の 2 つの値は 0.5 で重み付けして他の値と平均した。

5.2　結果

結果を表 3 に示す。CoLA や MRPC では破滅的忘却を防ぐ GS や EWC を用いたモデルが AdamW や SGD に比べ高いスコアを示した。EWC を用いて conj タスクを学習したモデルが最も高い平均スコア (72.22) を達成し、他のモデルも構文学習なしのスコア (70.54) を上回った。つまり多様な構文情報が性能向上に寄与し、中でも意味情報を保持しつつ構文情報を “ほどほどに” 学習したモデルが高い性能を発揮していると考えられる。

6　まとめ

本論文では事前学習済モデルに構文知識を取り入れるために、破滅的忘却を防ぐ最適化関数 GS と EWC を導入し、4 つの構文追加学習を比較した。モデルに元々含まれる意味情報と追加学習で補った構文知識を併せ持ったモデルは CoLA, RTE, MRPC タスクにおいて高い性能を発揮した。

適切な最適化関数を用いて知識をモデルに追加する本手法は、構文情報以外の事前学習にも適用できる。また作成したモデルは事前学習済モデルと同じ形式を持ち、応用タスクでの活用に適している。

参考文献

[1] Ali Basirat and Joakim Nivre. Syntactic nuclei in dependency parsing – a multilingual exploration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1376–1387, 2021.

[2] Emanuele Bugliarello and Naoaki Okazaki. Enhancing machine translation with dependency-aware self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1618–1627, 2020.

[3] Ethan A. Chi, John Hewitt, and Christopher D. Manning. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5564–5577, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[5] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, Vol. 3, No. 4, pp. 128–135, 1999.

[6] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. UCPhrase: Unsupervised Context-Aware Quality Phrase Tagging. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 478–486, 2021.

[7] Hiroshi Kanayama and Ran Iwamoto. How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4063–4073, 2020.

[8] Sudipta Kar, Giuseppe Castellucci, Simone Filice, Shervin Malmasi, and Oleg Rokhlenko. Preventing catastrophic forgetting in continual learning of new natural language tasks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 3137–3145, 2022.

[9] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, Vol. 114, No. 13, pp. 3521–3526, 2017.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[11] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

[12] Dmitry Nikolaev and Sebastian Pado. Word-order typology in multilingual BERT: A case study in subordinate-clause detection. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pp. 11–21, 2022.

[13] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, Vol. 22, pp. 400–407, 1951.

[14] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, 2020.

[15] Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

[16] Yuanhe Tian, Yan Song, and Fei Xia. Improving relation extraction through syntax-induced pre-training with dependency masking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1875–1886, 2022.

[17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

[18] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1405–1418, 2021.

[19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

[20] Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. Syntax-enhanced pre-trained model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5412–5422, 2021.

[21] Zhixian Yang and Xiaojun Wan. Dependency-based mixture language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 7758–7773, 2022.

[22] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 5824–5836, 2020.

[23] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9628–9635, 2020.

[24] Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. SG-Net: Syntax-Guided Machine Reading Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9636–9643, 2020.

[25] 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼, 浅原正幸, 乾健太郎. 予測の正確な言語モデルがヒトらしいとは限らない. 言語処理学会第 27 回年次大会予稿集, 2021.

[26] 鈴木潤, 全炳河, 賀沢秀人. ニューラル言語モデルの効率的な学習に向けた 代表データ集合の獲得. 言語処理学会第 28 回年次大会予稿集, 2022.