CSJ を用いた日本語話し言葉 BERT の作成

勝又智
株式会社レトリバ
satoru.katsumata@retrieva.jp

坂田大直
株式会社レトリバ
hiromasa.sakata@retrieva.jp

1 はじめに

現在，大規模なテキストデータを元に学習された事前学習モデルが多く公開されている．特に近年では BERT [1] を用いた研究が盛んに行われている．BERT とは，Transformer [2] の Encoder に対して，Masked Language Model（MLM）と Next Sentence Prediction（NSP）と呼ばれる事前学習を行ったものである．日本語については，Wikipedia や SNS データなどの書き言葉[1)]を用いて事前学習した BERT がいくつか公開されている．

しかしながら，発話の書き起こしといった，話し言葉に注目した日本語事前学習モデルは著者の知る限り存在していない．そこで，本研究では，発話の書き起こしデータである日本語話し言葉コーパス（Corpus of Spontaneous Japanese; CSJ）[3] を用いて話し言葉 BERT の作成を行った．具体的には，Wikipedia で学習された書き言葉 BERT を元に，次の 2 つの手法を用いて話し言葉 BERT を作成した．

1. 一部の層のみを話し言葉データで追加学習
2. 分野適応の手法を用いて追加学習

1 つめの手法は，書き言葉と話し言葉の差は主に統語的な側面にあるという仮説のもと，BERT の学習を行う．BERT における統語情報の学習がどの層で行われてるのかに関して，複数の観点から研究が行われている [4, 5, 6]．彼らの主張に基づいて，一部の統語的な面を学習していると考えられる層のみを追加学習することで，BERT の層の全てを学習するより話し言葉の統語情報をうまく学習できるのか検証した．2 つめの手法は，話し言葉 BERT を作成するために，書き言葉で学習した BERT に対して話し言葉データで分野適応を行う．具体的には Gururangan ら [7] の手法を参考に，CSJ 以外の発話の書き起こしデータを用意し，それらを用いて BERT を分野適応した．

また，作成した話し言葉 BERT を評価するため，本研究では評価タスクを作成した．CSJ には様々なアノテーションがされており，その中から本研究は係り受け解析と文境界推定，重要文抽出の 3 つの評価タスクを作成した．係り受け解析は，作成したモデルが統語的な情報を捉えられているか，文境界推定と重要文抽出は応用的なタスクに転用可能かどうか調べるために行う．

実験結果から，BERT の一部の層のみを学習する手法は統語的な面が強いタスクに有効であることがわかった．また，分野適応の手法を用いて追加学習することにより，書き言葉 BERT は話し言葉 BERT に分野適応可能なことがわかった．本研究で作成した話し言葉 BERT は公開予定である．

2 関連研究

BERT は単言語コーパスに対して MLM と NSP と呼ばれる 2 つの目的関数を用いて事前学習を行っている．Devlin ら [1] は事前学習した BERT に対して，最終的に解きたいタスクで fine-tuning を行うことで高い精度が得られることを報告している．入力のある単語 w_i に対して，いくつかの Transformer Encoder 層を重ね，最終的に得られた h_{w_i} を BERT の単語単位の出力として使用している．また，文頭に [CLS] を入れ，対応する出力 $h_{[CLS]}$ を文単位の出力として使用している．私たちの知る限り，日本語では Wikipedia で事前学習したモデル[2)]，SNS で学習したモデル[3)]，ビジネス記事で学習したモデル[4)]，様々な Web テキストで学習したモデル[5)] が存在している．本研究は，これらの書き言葉データではなく，発話

1） 本研究では，Web 上で投稿されたテキストを書き言葉，発話を元に書き起こされたテキストを話し言葉とする．そのため，発話を元にしているか不明瞭な SNS データは本研究では書き言葉として扱う．

2） https://github.com/cl-tohoku/bert-japanese, https://github.com/yoheikikuta/bert-japanese, https://alaginrc.nict.go.jp/nict-bert/index.html
3） https://github.com/hottolink/hottoSNS-bert
4） https://stockmark.co.jp/news/2019-04-08-1224
5） https://github.com/laboroai/Laboro-BERT-Japanese

表 1　CSJ 5-gram LM に対する PPL

	PPL w/ OOV	PPL w/o OOV
Wikipedia	3233	513
国会議事録	936	235

の書き起こしデータで BERT の事前学習を行った．

Tenney ら [4] は BERT の前半の層，つまり 12 層の場合は 1-6 層が統語的な構造を，後半の層が意味的な情報を扱っていることを報告している．一方で，Jawahar ら [5] や Hewitt and Manning [6] は BERT 12 層の内，それぞれ 8-9 層と 6-9 層が統語的な特徴を捉えていると報告している．本研究では，書き言葉と話し言葉の差は統語的な側面に存在すると考えた．上記の先行研究を踏まえ，統語的な特徴を捉えている層のみを追加学習することで話し言葉に向けた BERT の作成を試みた．

Gururangan ら [7] は BERT の派生モデルである RoBERTa [8] の分野適応のため，domain-adaptive pretraining（DAPT）と task-adaptive pretraining（TAPT）を提案した．DAPT は事前学習された RoBERTa に対して，解きたいタスクが属している分野のラベルなしデータを用いて追加で事前学習を行う手法である．TAPT は事前学習された RoBERTa に対して，解きたいタスクのラベルは使用せず，テキストデータのみを用いて追加で事前学習を行う手法である．本研究は，話し言葉 BERT 作成に向けて TAPT と DAPT を用いた．

3　話し言葉 BERT の学習

3.1　Layer-wise Pretraining

本研究では書き言葉と話し言葉の差が統語に存在するという仮説のもと，BERT の学習を行った．具体的には，BERT の一部の層が統語的な学習を行っていることを踏まえて，既存の 12 層の書き言葉 BERT の一部の層のみに対して， CSJ を用いた MLM と NSP の追加学習（layer-wise）を行った．先行研究 [4, 5, 6] を参考に，本研究では 1-6 層のみを学習した場合（layer-wise$_{[1-6]}$）と 8-9 層のみを学習した場合（layer-wise$_{[8-9]}$）と 6-9 層のみを学習した場合（layer-wise$_{[6-9]}$）を実験している．

3.2　DAPT/TAPT Pretraining

本研究では TAPT と DAPT を用いて，書き言葉で学習された BERT を話し言葉に分野適応する．

表 2　データサイズ

	サイズ	総単語数	講演数
国会議事録	11GB	2.3B	-
CSJ 全体	44MB	7.5M	3.3K
コアデータ	3MB	0.5M	0.2K
非コアデータ	41MB	7M	3.1K

TAPT については，CSJ を用いて評価を行うことから，CSJ のラベルなしデータを用いた MLM と NSP の追加学習を書き言葉 BERT に対して行っている．

本研究は DAPT に向けて，CSJ 以外の発話の書き起こしデータとして国会議事録[6] を用意し，MLM と NSP の追加学習を行っている．事前実験として，国会議事録は本当に CSJ に近しい分野なのかを確認した．具体的には CSJ データから 5-gram 言語モデル（LM）を作成，この LM を元に国会議事録のパープレキシティ（PPL）を計測した．また比較として CSJ LM に対する Wikipedia の PPL も計測している．この比較の際，国会議事録と Wikipedia のデータサイズはほとんど同じになるように調整している．表 1 が Out-of-Vocabulary（OOV）を含んだ場合，含まない場合の PPL である．この結果から，少なくとも Wikipedia と比べると国会議事録の方が CSJ に LM の観点では近しいことがわかる．

4　実験

本研究では，書き言葉 BERT に対して layer-wise または DAPT/TAPT を用いた話し言葉への分野適応を行った後に，話し言葉検証タスクを行った．

4.1　実験設定

本研究では書き言葉 BERT として Wikipedia で学習された BERT[7]（Wikipedia BERT）を使用した．また，CSJ コアデータを評価タスクに，コアデータと非コアデータ全てを BERT の追加学習に使用した．データサイズを表 2 に示す．単語分割は MeCab[8] と WordPiece[9] [9] を用いている．

layer-wise の学習は，CSJ 全てを用い，バッチサイズ 8 で 90K ステップ行っている．この時の最大文長は 512 としている．詳細は付録 A.1 に記載した．

DAPT には国会議事録を使用している．DAPT は

6）https://kokkai.ndl.go.jp
7）https://github.com/cl-tohoku/bert-japanese の bert-base-japanese-whole-word-masking を使用．
8）https://taku910.github.io/mecab
9）本論文では形態素と subword をまとめて**トークン**と呼称する．

表 3　話し言葉検証実験の結果

	係り受け解析		文境界推定			重要文抽出		
	UUAS	NED	Precision	Recall	F	Precision	Recall	F
Wikiepdia BERT	39.4	43.2	62.0	61.3	61.6	45.5	31.2	36.8
layer-wise$_{[1-6]}$	**44.6**	**47.9**	**63.6**	**66.1**	**64.8**	41.5	31.6	35.4
layer-wise$_{[8-9]}$	44.3	47.6	62.4	64.6	63.5	43.7	29.7	34.9
layer-wise$_{[6-9]}$	44.5	47.7	62.4	64.9	63.6	43.6	33.5	37.1
layer-wise$_{all}$	44.1	47.5	63.1	65.9	64.5	**44.0**	**33.8**	**37.5**
DAPT512	41.7	45.2	63.1	62.2	62.6	45.2	32.8	37.5
TAPT512	43.4	46.8	**64.5**	**64.1**	**64.3**	42.9	31.3	35.9
60K step						47.2	35.5	40.2
DAPT128-TAPT512	42.9	46.3	64.2	63.7	64.0	**45.6**	**35.4**	**39.7**
DAPT512-TAPT512	**43.7**	**47.0**	63.7	63.2	63.5	39.1	32.2	34.4

最大文長を 128 または 512 に設定，TAPT は 512 に設定している．DAPT は 最大文長に応じてバッチサイズや学習ステップ数を変更している．詳細は付録 A.2 に記載した．TAPT はバッチサイズ 8 で 20K ステップ行っている．本研究は追加学習として DAPT のみ，TAPT のみ，DAPT の後に TAPT（DAPT-TAPT）の 3 つのパターンを実験した．

4.2　話し言葉 BERT の検証実験

話し言葉 BERT の評価のため，CSJ コアデータから検証タスクを作成した．具体的には，コアデータを 5 等分し，3:1:1 の割合で学習，開発，テストデータを作成した．

係り受け解析　作成した話し言葉 BERT が統語的な面を捉えることができているか確認するため，係り受け解析を評価タスクの 1 つとした．BERT を用いた，教師なしでラベルなしの係り受け解析を行う手法 [10] によって，学習したモデルが CSJ にアノテーションされた係り受け構造を推定できるか実験を行った．各文節の分散表現は各トークンの分散表現を平均したものを使用している．テストデータに対して推論を行い，評価尺度は Undirected Unlabeled Attachment Score（UUAS）[11] と Neutral Edge Direction（NED）[12] を用いている．

実験の結果を表 3 に示す．この結果から，layer-wise と DAPT/TAPT どちらも Wikipedia BERT よりも高い精度であることがわかる．

layer-wise については，UUAS と NED のどちらの評価尺度でも layer-wise$_{[1-6]}$ が最も良かった．ブートストラップ検定を用いたところ，layer-wise$_{all}$ に対する layer-wise$_{[1-6]}$，layer-wise$_{[6-9]}$ の結果は，UUAS と NED どちらも統計的に有意（$p < 0.01$）であった．このことから，全ての層を学習するより，一部の層のみ，特に前半の層のみを学習した方が効果的であることがわかる．

DAPT/TAPT については，DAPT512-TAPT512 が最も精度が高かった．DAPT と TAPT で精度を比較すると，TAPT512 の方が高く，これらを組み合わせることで精度が向上することから，DAPT の学習は TAPT の学習に対してうまく補うことができていると考えられる．

文境界推定　書き言葉と違い，本研究が対象としている話し言葉には句点などの記号が含まれず，文境界が自明でない．そのため，話し言葉の応用的なタスクの 1 つとして，文境界推定を行う．具体的には CSJ の文境界のアノテーション[10] を IOB2 形式に変換し，BERT を用いて推定を行う．入力は 1 講演であり，この入力を 512 トークンごとに区切り，IOB ラベルを推定する．単語を subword に分割した場合，最初のトークンに対する推論をその単語に対する推論結果とする．評価は Precision，Recall，F-measure を使用する．

実験結果を表 3 に示す．この結果から，全ての手法は Wikipedia BERT より高い精度となっていることがわかる．

layer-wise 内で見てみると，layer-wise$_{[1-6]}$ は layer-wise$_{[all]}$ より Precision，Recall どちらも高い結果を示している．一方で，layer-wise$_{[8-9]}$ や layer-wise$_{[6-9]}$ は layer-wise$_{[all]}$ と比べて精度が低い．これらの結果から，少なくとも文境界推定というタスクにおいては前半の層のみを学習するのは有効だが，中間の層のみを学習するのは効果が薄いことがわかる．

10）CSJ には文境界ではなく，節境界としてアノテーションされている．

DAPT/TAPT については，TAPT512 が最も高い精度だった．TAPT512 に比べると DAPT512 の学習は精度が低いが，Wikipedia BERT と比べると高い結果であることがわかる．

重要文抽出　話し言葉の応用的なタスクの 1 つとして，重要文抽出を行う．CSJ には各文ごとに重要文かどうかの 2 値のアノテーションが振られている．入力 x が 1 講演で，x 内の各文頭に [CLS] トークンを付与し，$h_{[CLS]}$ に対して重要文かどうかの 2 値分類を推定する．本研究では Liu and Lapata [13] の BERT を利用した抽出型要約モデルを重要文分類に利用している．この検証タスクは 5 分割交差検証を用いており，評価は重要文かどうかについての 2 値分類について，重要文ラベルに対する Precision, Recall，F-measure を用いている．

実験結果を表 3 に示す．係り受け解析や文境界推定の結果と違い，重要文抽出においては全ての手法が Wikipedia BERT より高い結果とはならなかった．

layer-wise については layer-wise$_{[all]}$ と layer-wise$_{[6–9]}$ は Wikipedia BERT より高い結果になっているが，他の一部の層のみを学習したモデルは低い結果となっている．また，layer-wise$_{[all]}$ が layer-wise 内で最も高い精度であることから，重要文抽出では一部の層のみを学習する手法は効果が薄いことがわかる．

DAPT/TAPT については，係り受け解析や文境界推定の結果と異なり，DAPT512 は TAPT512 より精度が高い結果となった．また，最も精度が高かったのは DAPT128-TAPT512 だった．一方で，DAPT512-TAPT512 は，Wikipedia BERT よりも低い結果となっている．これらの結果から，重要文抽出においては DAPT と TAPT を組み合わせる手法は有効ではあるが，最大文長のパラメータによって結果が大きく変わることがわかる．

また，TAPT512 については，学習量が足らないと考え，TAPT の学習ステップを 20K ではなく，80K まで 20K 刻みで行った．詳細な結果は付録 B に示し，表 3 では最も高かった TAPT 60K ステップの結果を記載している．この結果から分かる通り，TAPT512 について，学習量を増やすことで Wikipedia BERT より高い結果となっている．

5　考察

文境界推定は単語の活用，特に終止形を当てる必要があると考えられ，統語を捉える必要があるタスクであると考えられる．この文境界推定や係り受け解析に対しては一部の層のみを学習することの有効性が確認できた．特に，前半の層のみを学習した場合は全ての層を学習するより高い精度を出している．一方で，文の意味を捉える必要があるタスクである，重要文抽出では一部の層のみを学習するアプローチは効果が薄い．これは，一部の層のみを学習するアプローチは統語的な情報は捉えることができるが，意味的な面ではうまく捉えることができていないからだと考えられる．仮に，Teeny ら [4] の主張の通り，前半の層が統語を，後半の層が意味的な情報を扱っていると考えた場合，後半の層をうまく学習することができれば重要文抽出でも精度が向上すると考えられる．

係り受け解析や文境界推定では DAPT よりは TAPT の方が効果があったが，重要文抽出では DAPT の方が有効であり，TAPT は学習量を増やすことで DAPT と TAPT を組み合わせたものと同じくらいの結果となった．これらの結果から，統語を捉える必要があるタスクは TAPT だけでも十分ではあるが，意味を捉える必要があるタスクだと TAPT はより多く学習を行わなければいけないことがわかる．一方で，DAPT と TAPT を組み合わせることでどのタスクでも精度向上が見込める．これは，DAPT と TAPT を連続して行うことで話し言葉分野の様々なデータを学習に使用し，タスクに対して頑健なモデルができるからではないかと考える．

6　おわりに

本研究では，話し言葉 BERT 作成に向けて，一部の層のみの学習と分野適応の 2 つの手法を用いた．また，CSJ から係り受け解析，文境界推定，重要文抽出といった BERT を用いた 3 つの評価タスクを設計した．実験の結果，BERT の一部の層のみを学習する手法は統語的なタスクに有効であることがわかり，簡単な分野適応手法で話し言葉モデルを作成できることがわかった．作成したモデルはそれぞれ**日本語話し言葉部分学習 BERT**，**日本語話し言葉分野適応 BERT** として公開予定である．

謝辞

この研究は，国立国語研究所との共同研究で行ったものです．国立国語研究所の浅原正幸様，前川喜久雄様，小磯花絵様には有益な助言をいただき，岡照晃様には研究環境管理などのご支援賜りました．この場を借りて深く御礼申し上げます．

参考文献

[1]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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ¥ L ukaszKaiser, Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 第30巻, pp. 5998–6008. Curran Associates, Inc., 2017.

[3]Kikuo Maekawa. Corpus of spontaneous japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[4]Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[5]Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[6]John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. 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. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7]Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics.

[8]Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, Vol. abs/1907.11692, , 2019.

[9]Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152. IEEE, 2012.

[10]Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4166–4176, Online, July 2020. Association for Computational Linguistics.

[11]Dan Klein and Christopher Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 478–485, Barcelona, Spain, July 2004.

[12]Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 663–672, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[13]Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. 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. 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics.

A　詳細な実験設定

A.1　layer-wise の詳細な実験設定

layer-wise のハイパーパラメータを表 4 に示す.

表 4　layer-wise のハイパーパラメータ

パラメータ名	値
最大文長	512
学習ステップ数	90K
バッチサイズ	8
学習率	5e-5
GPU	NVIDIA Tesla T4 1 枚

A.2　DAPT/TAPT の詳細な実験設定

DAPT/TAPT のハイパーパラメータを表 5 に示す.

表 5　DAPT/TAPT のハイパーパラメータ

	DAPT		TAPT
最大文長	128	512	512
学習ステップ数	10K	30K	20K
バッチサイズ	32	8	8
学習率	1e-4		
GPU	NVIDIA Tesla T4 2 枚		

B　TAPT の学習量を増やした際の重要文抽出の結果

TAPT のみの学習ステップ数が{20K, 40K, 60K, 80K}の時のモデルを重要文抽出に使用した結果を表 6 に示す.

表 6　TAPT の学習量を蓋した際の重要文抽出結果

	Precision	Recall	F
Wikipedia BERT	45.5	31.2	36.8
TAPT 20K	42.9	31.3	35.9
TAPT 40K	44.5	**37.2**	40.1
TAPT 60K	**47.2**	35.5	**40.2**
TAPT 80K	47.1	34.7	39.3