BERT はどのように逆接の談話関係を判定しているか ― Attention と品詞を手がかりとして―

佐藤拓真 [1]　窪田愛 [2]　峯島宏次 [1]
[1] 慶應義塾大学 [2] 東京大学
takuma1229@keio.jp aikubota@g.ecc.u-tokyo.ac.jp
minesima@abelard.flet.keio.ac.jp

概要

本研究は、BERT が自然言語処理・計算言語学の中でも特に複雑な意味構造の把握が必要な逆接等の談話関係認識を行う際に、どのような品詞に Attention を向けているかを明らかにする。分析の結果、BERT は談話関係認識において、各品詞に満遍なく注意を向けているものの、各 layer や head はそれぞれ異なる品詞に注意を向けており、アーキテクチャにおいてそれぞれの layer や head が異なる役割を果たしているという先行研究の結果が日本語における逆接等の談話関係認識においても同様に認められることが明らかになった。

1　はじめに

自然言語には、「また」や「しかし」といった、文や節の間にある意味的な関係・つながりを示す様々な表現が存在し、言語学・自然言語処理いずれの分野でも研究の対象となっている [1, 2]。この意味的な関係は談話関係 (discourse relation) と呼ばれ、談話関係がどのように決定されるかを解明することは、言語理解において意味的なつながりや構造がどのように把握されているかを解明することに繋がる。

本研究では、Transformer ベースの言語モデルである BERT [3] が逆接等の談話関係を判定する際にどのような品詞に注意を向けているかを、モデルの推論時の Attention Weight を分析することで明らかにする。また、BERT の各 layer や head がそれぞれ異なる言語的側面を捉えているという先行研究における知見 [4, 5] が、日本語においても同様に認められるか、さらに、複雑な意味構造の把握を必要とする逆接等の談話関係認識においても認められるかを明らかにする。

これらの問いに答えるために、本研究では、逆接等の談話関係認識に特化したデータセットで BERT を fine-tuning して談話関係認識モデルを構築したあと、test データに対するモデルの推論において、分類問題における予測を総合的に表現するとされる [CLS] トークンからの Attention Weight を分析することで、モデルがどのような品詞に注意を向けて談話関係認識を行っているかを検証する（図 1）。その際に、モデルの各 layer や head ごとにも分析を行う。

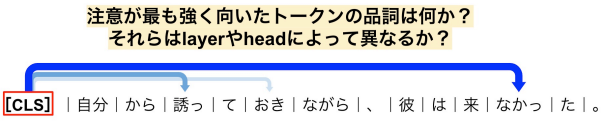

図 1　分析の概略図

分析の結果、BERT は日本語談話関係認識において、全体としては入力トークンの品詞に一定程度満遍なく注意を向けていることが明らかになった。また、英語を対象とした先行研究が示すのと同様に、各 layer や head が最も強い注意を向けている品詞はそれぞれ異なり、日本語談話関係認識においてもアーキテクチャにおける個々の要素がそれぞれ異なる役割を果たしていることが示唆された。本研究の結果は、自然言語処理や計算言語学の視点から、言語における意味的な構造の把握のなかでも、特に逆接がもつ暗黙の推論関係がどのように認識されるかの解明に寄与するものとして位置づけられる。

2　背景と関連研究

談話関係 (discourse relation) とは、文や節の間にある意味的な関係・つながりのことであり [6, 7]、自然言語処理の領域において計算機を用いて談話関係を分類するタスクは、談話関係認識 (discourse relation recognition) と呼ばれる [8]。日本語の談話関係についてアノテーションを行った先行研究は複数存在する [6, 7, 9]。この内、窪田ら (2023) によるもの [9] は、表層的特徴としては表れない談話関係における複雑

な意味論的構造に焦点を当てるものであるため、本研究ではこれをデータセットとして採用した。このデータセットは、接続詞「ながら」「つつ」を対象として談話関係ラベルのアノテーションを行ったものである。これらの接続詞は「同時進行」と「逆接」の異なる二つの談話関係を示しうるが、以下の例文のようなケースでは、この二つの談話関係のどちらが正しいかを判別することが容易でない。

(1) あなたは私をおだてながら、内心はお人好しのバカだと思ってるんでしょ。

窪田ら (2023) は、逆接の談話関係が以下の例文のように暗黙的な推論を伴うことに着目してアノテーションを行なっている [9]。

(2) a. 自分から誘っておきながら、彼は来なかった。（逆接）
b. 自分から誘ったならば、普通はきちんと来るだろう。（暗黙の推論）

他方、Transformer ベースの言語モデルが著しい発展を見せる中で、これらのモデルが内部的にどのようなメカニズムで動作しているかを説明する手法が盛んに探求されている [10]。それらの手法の一つに、モデルの Attention を分析することで、モデルがいかなる特徴に「注意」を向けているかを説明するものがある [11, 12]。Attention の分析を行った先行研究においては、Transformer ベースの言語モデルのの Attention が一定の構文情報や意味情報といった知識を獲得しており、さらに各 layer や head がそれぞれ異なる言語的知識を獲得していることが示唆されている [4, 5, 13, 14, 15]。例えば、Tenney ら (2019)[14] はモデルの浅い層は構文情報を、深い層はより複雑な意味構造を捉えていることを、Vig ら（2019）[13] は個々の head がそれぞれ特定の品詞に強い注意を向けていることを報告している。

しかしながら、Transformer ベースの談話関係認識モデルがどのような根拠によって談話関係を判断しているのか、その仕組みは解明されていない。したがって、本研究は、そのようなモデルの Attention を分析することで、その仕組みの解明に寄与する。

3 談話関係認識モデルの構築

3.1 タスクの定式化

本研究で行う談話関係認識タスクは、入力トークン列 $S = \{w_1, \ldots, w_d\}$ に対して正しい談話関係ラベル $L \in \{l_1, \ldots l_n\}$ を 1 つ判定することを目的とする、多クラス分類タスクとみなすことができる（図 2）。ただし、w_i はトークン列における i 番目のトークン、d はトークン列の長さ、l は談話関係ラベル、n は文書集合における談話関係ラベルの数を表す。

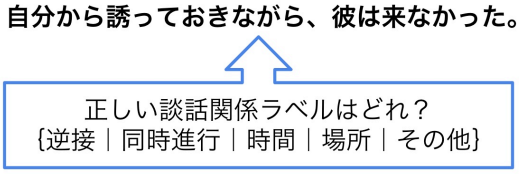

図 2 談話関係認識タスク

表 1 データセット

表現	Kappa 係数	分類ラベル	件数
ながら	0.72	逆接	213
		同時進行	1047
		その他	65
つつ	0.46	逆接	51
		同時進行	186
ところで	0.75	逆接	27
		時間	14
		場所	49
		その他	18

3.2 データセット

使用したデータセットは、接続詞「ながら」「つつ」を含む文書を対象とした窪田らによる日本語談話関係データセット [9] に、新たに接続詞「ところで」を含む文集合に対するアノテーションを行い追加したものである。このデータセットの文は、文に対する品詞タグと統語構造が付与されている Kainoki Treebank[16] から、複雑な統語的条件のもと抽出したものである。データセットを 8:1:1 の割合で分割し、それぞれ train / valid / test データとした。

なお、結果の解釈性向上の観点から、談話関係ラベルは [9] において設定されているものに少々の簡略化を加えた[1)]。簡略化後の接続詞と各談話関係ラベルの件数の対応、およびアノテーションにおける Kappa 係数を表 1 に示す。なお、「その他」のラベルは、イディオムや慣用的な表現のほか、構文的に談話として捉えることが困難なものなどといった例外的なケースに付与されたものである。

1) 談話関係ラベルと例文を、Appendix の表 A に示す。

表 2　fine-tuning の結果（test データでの実験）

	Prec	Rec	F1	正解率	件数
逆接	0.85	0.73	0.79	-	30
同時進行	0.92	0.96	0.94	-	128
時間	0.67	1.00	0.80	-	2
場所	1.00	0.97	0.80	-	3
その他	0.75	0.60	0.67	-	5
全体	0.83	0.79	0.79	0.90	168

3.3　BERT の fine-tuning による談話関係認識モデルの構築

上述のタスク設定とデータセットのもと、Transformer ベースの Encoder-only モデルである BERT [3] を fine-tuning して、談話関係認識モデルを構築した[2]。事前学習済みモデルおよびトークナイザとして、Hugging Face 上で公開されている cl-tohoku/bert-base-japanese-v3[3] を使用した。モデルは 12 層の layer（隠れ層）と 12 個の head をもつ。

fine-tuning 後のモデルを用いて test データで予測を行った際の結果を表 2 に示す。なお、データセットのクラスは不均衡に分布しているため、全体における各スコアは Macro 平均を用いて算出した。BERT を用いた日本語談話関係認識タスクの先行研究 [17] における結果や、データセットに対する談話関係ラベル付与アノテーション時のカッパ係数を考慮し、本研究において、ここで fine-tuning を行った談話関係認識モデルの推論における Attention 分析を行うことの有意味性は十分に高いと判断した。本タスクのデータセットにおいて、談話関係の判別が接続表現だけで決定されず、文脈や常識的な知識などを参照する必要があることを考慮すると、BERT が本タスクで表 2 のような高い能力を示していることは、直観に反する興味深い結果である。

4　推論時の Attention の分析

4.1　分析の概要

上述のモデルを用いて test データにおいて談話関係認識を行った際の、推論時の Attention Weight の分析を行った。具体的には、入力となるトークン列において、[CLS] トークンの Attention Weight を取得し、[CLS] トークンから最も強く注意を向けられているトークンの品詞が何であるかを分析した (図 1)。

2）　学習時のハイパーパラメータを Appendix の表 B に示す。

3）　https://huggingface.co/cl-tohoku/bert-base-japanese-v3, 2024 年 1 月 2 日閲覧

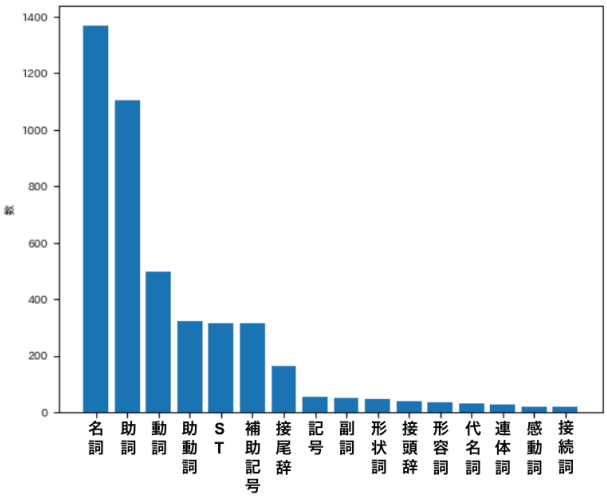

図 3　分析対象の文書集合の品詞分布

[CLS] トークンは分類問題において集約された文レベルの表現であり、分類問題におけるモデルの予測を総合的に表現するものと考えられている [5] ため、これを Attention 分析の対象トークンとした。

実装の都合上、test データの 168 文中、言語モデルにおけるトークナイズと MeCab による形態素解析時の分かち書きのパターンを完全に一致させることが可能である 100 文のみを分析の対象とした。分析対象の 100 文中に出現する品詞の分布を図 3 に示す。なお、BERT の注意は “[SEP]” などの special token に強く向けられる傾向があるが、これらのトークンへの注意は出力にほぼ影響を与えない “no-op” である [4] ため、special token と日本語における句読点を stop token (ST) として分析から除外した。

4.2　分析の詳細と結果

上述の方法のもと、Attention Weight の分析を 3 パターン行った。

4.2.1　全 layer/head の Attention の総和の分析

1 つ目のパターンとして、推論時のモデルにおける全ての layer と head の Attention Weight の総和を算出し、分析を行った。モデルの n 層目の layer, m 番目の head における Attention Weight ベクトルを $\boldsymbol{a}_{n,m} \in \mathbb{R}^d$ とする[4]。ここで、[CLS] トークンから最も強い注意を向けられたトークンのインデックスを MAX_ATT_IDX とすると、$\mathrm{MAX\_ATT\_IDX} = \underset{i \in \{1,\ldots,d\}}{\operatorname{argmax}} \sum_{n=1}^{12} \sum_{m=1}^{12} \boldsymbol{a}_{n,m}$ である。

4）　3.1 において述べた通り、d は入力トークン列の長さである。また、3.3 にて述べたモデル構造の設定より、本研究においては $n, m \in \{1, \ldots, 12\}$ となる

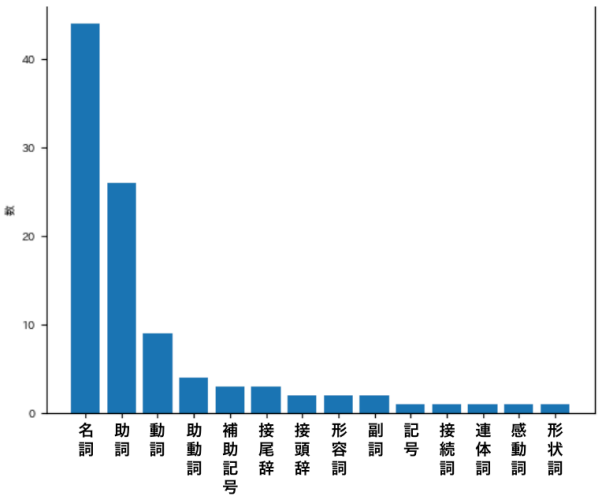

図 4　最も強い注意を向けられたトークンの品詞分布

このときの $w_{\text{MAX\_ATT\_IDX}}$ の品詞の分布を図 4 に示す。図 3 と図 4 を比較すると、特に上位の品詞では順位や分布の仕方が一定程度類似しているため、BERT は全体としては、概ね全ての品詞に満遍なく注意を向けて分類を行っていることが示唆される。

4.2.2　layer ごとの attention 分析

2 つ目のパターンとして、各 layer における全 head の Attention Weight の総和を算出し、分析を行った。このとき、m 番目の head における Attention Weight ベクトルを $\boldsymbol{a}_m \in \mathbb{R}^d$ とし、その他はすべて 4.2.1 と同様の記号法のもとで、各 layer について $\text{MAX\_ATT\_IDX} = \underset{i\in\{1,\ldots,d\}}{\text{argmax}} \sum_{m=1}^{12} \boldsymbol{a}_m$ である。

各 layer において $w_{\text{MAX\_ATT\_IDX}}$ の品詞として最も多かったものを、表 3 に示す。各 layer において最も強い注意を頻繁に向けられた品詞が異なることは、日本語談話関係認識においても、先行研究が指摘する [5, 13] ように各 layer がそれぞれ異なる役割を果たしていることを示している。BERT の浅い層は構文情報を、深い層はより複雑な意味的情報をそれぞれ処理しているとする先行研究 [14] と本研究の結果から、談話関係認識における意味的情報の担い手となる要素が何であるかを分析することが、今後の研究方針として考えられる。

4.2.3　head ごとの attention 分析

3 つ目のパターンとして、各 head における全 layer の Attention Weight の総和を算出し、分析を行った。このとき、n 層目の layer における Attention Weight ベクトルを $\boldsymbol{a}_n \in \mathbb{R}^d$ とし、その他はすべて

表 3　各 layer において最も強い注意を向けられたトークンの品詞（最上位のもの）

layer	1	2	3	4
品詞	助詞	助動詞	助動詞	助動詞
layer	5	6	7	8
品詞	助動詞	助動詞	助動詞	助動詞
layer	9	10	11	12
品詞	助動詞	動詞	助動詞	名詞

表 4　各 head において最も強い注意を向けられたトークンの品詞（最上位のもの）

head	1	2	3	4	5	6
品詞	助詞	助詞	助詞	助詞	名詞	名詞
head	7	8	9	10	11	12
品詞	名詞	名詞	助詞	名詞	助詞	名詞

4.2.1 と同様の記号法のもとで、各 head について $\text{MAX\_ATT\_IDX} = \underset{i\in\{1,\ldots,d\}}{\text{argmax}} \sum_{n=1}^{12} \boldsymbol{a}_n$ である。

各 head において $w_{\text{MAX\_ATT\_IDX}}$ の品詞として最も多かったものを、表 4 に示す。head ごとに最も注意が強く向く傾向のある品詞が異なることは、先行研究の結果 [13] と一致するものである。各 head が談話関係認識においていかなる「分業」を行っているかを他の先行研究 [4, 5, 15] を参考に明らかにすることが、今後の研究方針として考えられる。

5　おわりに

本研究では、日本語談話関係認識タスクにおける BERT の推論時の Attention を分析し、BERT は談話関係認識において各品詞にまんべんなく注意を向けているが、各 layer と head はそれぞれ異なる品詞に注意を向けていることを明らかにした。

しかしながら、本研究の結果から、逆接の談話関係認識における BERT の予測根拠が完全に明らかになった訳ではない。したがって、今後の課題として、BERT の Attention が強く向いている先のトークンとの距離や、品詞以外の言語的情報と Attention の関係について分析を行うことで、Transformer ベースの言語モデルが逆接のような複雑な意味関係をいかなる仕方で捉えているかをさらに調査することなどが挙げられる。本研究が、言語モデルや人間の意味理解について、自然言語処理・計算言語学からのアプローチとして有意味な一部分となることを期待する。

謝辞

本研究は、JST CREST、JP-MJCR2114 の支援を受けたものである。

参考文献

[1] Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. Implicit discourse relation classification via Multi-Task neural networks. **AAAI**, Vol. 30, No. 1, 2016.

[2] Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. Where are we in discourse relation recognition? In **Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 314–325. Association for Computational Linguistics, 2021.

[3] 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. Association for Computational Linguistics, 2019.

[4] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286. Association for Computational Linguistics, 2019.

[5] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 842–866, 2020.

[6] 岸本裕大, 村脇有吾, 河原大輔, 黒橋禎夫. 日本語談話関係解析：タスク設計・談話標識の自動認識・コーパスアノテーション. 自然言語処理, Vol. 27, No. 4, pp. 889–931, 2020.

[7] 金子貴美. 日本語談話関係認識のための理論とコーパス構築. PhD thesis, お茶の水女子大学大学院人間文化創成科学研究科理学専攻, 2020.

[8] Wei Xiang and Bang Wang. A survey of implicit discourse relation recognition. **ACM Comput. Surv.**, Vol. 55, No. 12, 2023.

[9] 窪田愛, 佐藤拓真, 天本貴之, 秋吉亮太, 峯島宏次. 逆接の推論関係に着目した日本語談話関係アノテーション. 言語処理学会第 29 回年次大会 発表論文集, pp. 375–380, 2023.

[10] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. 2023.

[11] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. **IEEE Transactions on Neural Networks and Learning Systems**, Vol. 32, No. 10, p. 4291–4308, 2021.

[12] Xiaobing Sun and Wei Lu. Understanding attention for text classification. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3418–3428. Association for Computational Linguistics, 2020.

[13] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 63–76. Association for Computational Linguistics, 2019.

[14] 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. Association for Computational Linguistics, 2019.

[15] Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. Do attention heads in BERT track syntactic dependencies? **CoRR**, Vol. abs/1911.12246, , 2019.

[16] The Kainoki Treebank – a parsed corpus of contemporary Japanese,（2022 年 1 月 9 日閲覧）. https://kainoki.github.io.

[17] Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 1152–1158. European Language Resources Association, 2020.

付録（Appendix）

表 A: 談話関係ラベルと例文

表現	分類ラベル	例文
ながら	逆接	自分から誘っておきながら、彼は来なかった
	同時進行	音楽を聞きながら勉強をしていた。
	その他	残念ながら、不合格だった。
つつ	逆接	悪いと思いつつ、ついやってしまう。
	同時進行	各所の意見を踏まえつつ、対策を講じます。
ところで	逆接	努力したところで、どうせ報われないさ。
	時間	私が席についたところで、教授が姿を見せた。
	場所	遠いところで誰かの声が聞こえる。
	その他	私の聞いたところでは、彼女は引っ越すらしい。

表 B: ハイパーパラメータ

パラメータ	値
epochs	15
loss function	Cross Entropy Loss
batch size	16
word tokenizer type	mecab
embedding dimension	768
optimizer	AdamW
learning rate	5e-05
weight decay	0.01
dropout	0.1