予測根拠として解釈性の高いアテンションの選択

石井 愛　　小松 祐城　　脇森 浩志
日本ユニシス株式会社
{ai.ishii, yuki.komatsu, hiroshi.wakimori}@unisys.co.jp

1 はじめに

高い精度で予測を行うことができる深層学習を実サービスに適用していくにあたり，モデルの解釈性への関心が高まっている．自然言語処理において重要な進歩をもたらしたアテンション [1] は，その重みにより特定の単語にどれだけ重みを付けたかを知ることができるため，アテンションを用いてブラックボックスであるモデルを解釈しようとする研究が盛んである．

文書分類タスクにおいては，アテンションの重みを変更しても予測精度に大きく影響がないことから，アテンションの解釈性を否定する研究がある [2, 3]．しかし，セルフアテンションに基づく Transformer[4] を用いた BERT[5] では，文書分類タスクにおいてもアテンションの重みの変更が予測精度に大きな影響を与えることが示されている [6]．ただし，この研究において，セルフアテンションの重みの解釈性について，人間が根拠とする箇所との一致度については評価されていない．

また，BERT について，個々のアテンションヘッドの予測における重要性は異なること [7, 8] や，異なる文法的な役割を持つ [9] ことが示されている．そこで，予測の解釈性においても個々のアテンションヘッドの重要性は異なるのではないかと考えた．

本稿では，多くのタスクで優れた性能を達成した BERT の個々のアテンションヘッドに焦点を当て，文書分類の予測の解釈性について，人間が根拠とする箇所との一致度に基づく定量的な評価指標 (2.1 節) と，それを用いたアテンションの選択手法 (2.2 節) を提案する．評価指標およびアテンションの選択について，複数のデータセットを用いて提案手法の有効性を示す (4.1 節，4.2 節)．さらに，手作業による定性的な評価 (4.3 節) により，有効性を裏付ける．

本研究の貢献は次のようにまとめられる．(i) アテンションの重みの解釈性を定量化した指標により，解釈性が相対的に高いアテンションヘッドを選択できることを示した．解釈性の高いアテンションヘッドは，最終層 (12 層目) のみではなく 10〜12 層目に存在し，予測クラスにより異なることがわかった．(ii) 提案手法は，アテンションの可視化による予測根拠提示における解釈性の改善に寄与することができる．

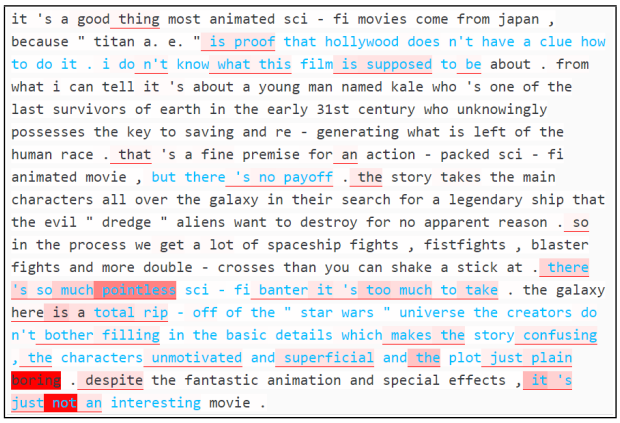

図 1: Movie Reviews のサンプル．水色文字：正解データの根拠箇所，文字背景の赤色濃さ：アテンションの強さ，下線：上位 20%のアテンション

2 提案手法

2.1 解釈性の評価指標

根拠としての妥当性を評価するベンチマークとして，ERASER ベンチマーク [1] が提案された [10]．ERASER では根拠個所がアノテートされた複数種類のデータセットおよびその評価指標が提案されている．文書分類のデータセットである Movie Reviews は，映画のレビューにポジティブ，ネガティブのクラスと，その根拠となる箇所がアノテートされたデータセット [11] を再整備したものである．Movie Reviews データセットのサンプルを図 1 に示す．

ERASER の評価指標のうち，Soft タイプとされている以下の指標により，アテンションの重みなどの連

[1] http://www.eraserbenchmark.com

データセット		根拠数	トークン数	根拠のトークン比率	文書数
オリジナル	train	8.7	773.6	66.8 (9.4%)	1600
	dev	7.6	761.5	49.8 (7.2%)	200
	test	10.4	795.0	240.8 (31.4%)	199
BERT 用	train	4.7	446.7	36.4 (8.3%)	1600
	dev	4.2	442.3	28.3 (6.6%)	200
	test	5.5	451.3	111.1 (25.2%)	199

表 1: データセットの統計 (Movie Reviews)

データセット		トークン数	文書数
books	train	170.4	1500
	dev	190.6	100
	test	177.1	400
dvd	train	165.3	1500
	dev	184.6	100
	test	167.7	400
livedoor	train	668.7	4421
	dev	665.0	1473
	test	652.6	1473

表 2: データセットの統計 (その他)

データセット	バッチサイズ	エポック数	学習率
Movie Reviews	24	7	2e-5
Books	24	5	5e-5
DVD	24	6	5e-5
Livedoor	24	8	3e-5

表 3: BERT の各パラメータ

続的なスコアを評価することができる．

AUPRC: 根拠の正解とスコアの，Precision-Recall 曲線の作る面積．0 から 1 の範囲をとり，値が大きいほど良い．

Comprehensiveness: 根拠の箇所を削除した場合に，削除しない場合と比較し，予測の信頼性が低下が大きいほど根拠の網羅性が高いとする指標．文書 d_i のトークン x_i から，根拠箇所のトークン r_i を削除した根拠箇所の上位 k_d のトークンを削除した $\tilde{x}_i$ を作成する．文書分類において，$\hat{p}_{ij}$ をモデル m によるクラス j のオリジナルの予測確率: $\hat{p}_{ij} = m(x_i)_j$ とし，トークンを削除した $\tilde{x}_i$ に対する予測確率を，$\tilde{p}_{ij} = m(\tilde{x}_i)$ とする．

$$\text{comprehensiveness}_i = \hat{p}_{ij} - \tilde{p}_{ij} \tag{1}$$

Sufficiency: 根拠の箇所以外を削除した場合に，削除しない場合と比較し，予測の信頼性の変化が小さいほど根拠の充足性が高いとする指標．根拠箇所のトークン r_i のみで予測したクラス j の予測確率を $\bar{p}_{ij}$ とする．

$$\text{sufficiency}_i = \hat{p}_{ij} - \bar{p}_{ij} \tag{2}$$

ここで，根拠箇所のトークン r_i は，アテンションの上位 k_d のトークン群とする．

本研究では，アテンションの重みの解釈性の指標として，comprehensiveness と sufficiency から $\mathbf{I_{aw}}$(interpretability of attention weights) を設定する．comprehensiveness は大きいほうが良く，sufficiency は小さいほうが良い指標のため comprehensiveness から sufficiency を減算する．文書数を n とすると，

$$I_{aw} = \frac{1}{n}\sum_{i=1}^{n}(\bar{p}_{ij} - \tilde{p}_{ij}). \tag{3}$$

2.2 解釈性の高いアテンションの選択

複数のデータセットにおいて，BERT で予測を行う際の全 144 パターン (12 層，12 ヘッド) のアテンションの重みを出力し，I_{aw} を比較する．さらに，予測クラスによって，重要なアテンションヘッドが異なる可能性があるため，予測クラスごとに I_{aw} が最高値となるアテンションヘッドの層・ヘッドを調査し，以下のようにマージして α^{merge} を求める．

文書 d_i の予測がクラス j のとき，

$$\alpha_i^{merge} = \alpha_i^{TOP(d_j)}. \tag{4}$$

ここで，d_j は予測がクラス j となる文書群，$\alpha^{TOP(d_j)}$ は d_j について I_{aw} が最高値となるアテンションヘッドの重みである．

3 実験設定

直接根拠の正解との一致率を図るデータセットとして，Movie Reviews を使用する．Movie Reviews は BERT の 512 トークンの制限に合わせて文書および根拠箇所を切り取った (表 1)．根拠個所の正解データセットの無い通常の文書分類データセットとしては，レビューデータ (Books，DVD)[12] および Livedoor ニュースコーパス [2] を使用する (表 2)．

モデルの実装には，Transformers ライブラリ [3] の BertForSequenceClassification クラスを用いる．事前学習モデルは bert-base-uncased を用いる．Livedoor ニュースコーパスについては，日本語の事前学習モデル bert-base-japanese-whole-word-masking[4] を用いる．両事前学習モデルは，12 層，隠れ層 768 次元，12 ヘッド，110M パラメータのモデルである．

BERT の各パラメータは，著者らによって提唱されている組み合わせから，バッチサイズ∈ {16, 24} とエ

[2] https://www.rondhuit.com/download.html#ldcc
[3] https://github.com/huggingface/transformers
[4] https://github.com/cl-tohoku/bert-japanese

ポック数∈ {1, ..., 10} のみ変更し，グリッドサーチにより dev の精度が高い組み合わせを選択する (表 3)．できるだけ多くのトークンに対するアテンションの重みを検証するため，max sequence length は BERT に設定できる最大値である 512 で固定する．

アテンションは，各データセットの test を用いた予測の実行時に，すべての層・ヘッドの先頭の [CLS] トークンのアテンションの重みを取得する．アテンションの重みには min max 正規化を適用する．

4 実験結果

4.1 解釈性の指標の妥当性評価

図 2 に，Movie Reviews における，12 層 12 ヘッドの 144 パターンのアテンションの AUPRC と各指標のピアソン積率相関係数を示す．comprehensiveness は AUPRC と正の強い相関 (0.90)，sufficiency は負の強い相関 (-0.78) があり，I_{aw} ($k_d = 0.2$)[5] はさらに強い正の相関 (0.92) を示し，根拠としてのアテンションの選択に有用であることが示された．

また，層の番号と AUPRC との相関は 0.83 であり，深い層になるほど AUPRC が高くなる傾向があることがわかった．ただし，図 2 のグラフの点を層の番号で色分けして示したように，最終層に近い層では解釈性の値のばらつきが大きい．そのため，アテンションの重みの可視化で用いられることが多い最終層のアテンション重みの平均では，解釈性が低くなる可能性があることがわかった．

4.2 解釈性の定量評価

Movie Review において，予測クラス $\in \{NEG, POS\}$ ごとに I_{aw} を算出した結果を表 4 に示す．NEG では，11-2，POS では 11-11 が最高値となり，予測クラスによって I_{aw} が最高値となるアテンションヘッドが異なった．次に，各アテンションヘッドの I_{aw} および AUPRC を表 5 に示す．ランクとして I_{aw} が大きい順を示す．I_{aw} と AUPRC のランクは若干の入れ替わりが見られるが，上位数件に関しては，I_{aw} によって AUPRC の大きさを捉えることができている．差は小さいもの

[5] k_d は，表 1 の test における根拠箇所のトークン数の割合が 25.2%であることから，$k_d \in \{0.1, 0.2, 0.25\}$ の中で AUPRC と I_{aw} の相関を比較し，最大となった 0.2 を設定した．このパラメータはデータセットにより変動する可能性があるが，他のデータセットにおいても根拠個所のトークンの割合は同程度であると仮定し，本稿では他のデータセットにおいてもこのパラメータを使用する．

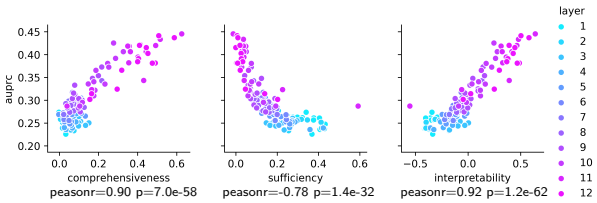

図 2: AUPRC と各指標の相関 ($k_d = 0.2$)

予測クラス	ランク	層-ヘッド	acc	I_{aw}
NEG	1	11-2	0.922	**0.692**
	2	11-11		0.687
	3	11-10		0.666
POS	1	11-11	0.948	**0.587**
	2	12-10		0.502
	3	11-3		0.477

表 4: 予測クラスごとの I_{aw}(Movie Reviews)

の，予測クラスごとにマージしたアテンションの重み α^{merge} の I_{aw} および AUPRC が最高値となった．

次に，根拠のアノテーションが無い文書分類データセットにおける調査結果を表 6 に示す．α^{merge} の算出に用いた層・ヘッドの肩に載せて，予測クラスの略称を示す．Movie Reviews と同様に，すべてのデータセットにおいて，α^{merge} の I_{aw} が一番大きい結果となった．よく可視化に用いられる最終層のアテンションの重みの平均（12-avg）は上位からランク $r = 7$〜20 番目の結果であり，より解釈性の高いアテンションヘッドが存在することが示された．

また，同じ事前学習モデルを用いた，Movie Reviews，Books，DVD においては，I_{aw} が上位となる層・ヘッドに共通のものが複数あることがわかった．これは，ドメインの異なる同じタスクのデータセット間で個々のアテンションヘッドの重要性が共通することを示した Michel ら [8] と同様の結果である．Michel らによると，その重要性はファインチューニング時の早い段階で決定されることから，事前学習の影響が大きいことが示唆される．事前学習で決定された個々のアテンションの文法的役割 [9] によって，同様のタスクで重要となるアテンションヘッドは共通するものになると考えられる．

4.3 解釈性の定性評価

I_{aw} が人間にとっての解釈性を評価できているかを調査するため，I_{aw} が最高値となった α^{merge} と，最終層の平均（12-avg）の 2 つを人手で評価する簡易的な定性評価を実施した．実験には Livedoor ニュースコーパスを用い，test の中の 100 個のサンプルについて，

ランク	層-ヘッド	I_{aw}	AUPRC
1	α^{merge}	**0.641**	**0.446**
2	$11-11^{POS}$	0.639	0.445
3	$12-10$	0.579	0.437
4	$12-4$	0.502	0.442
5	$11-2^{NEG}$	0.487	0.434
6	$12-9$	0.479	0.381
7	$11-1$	0.467	0.42
8	$12-avg$	0.464	0.381
9	$12-5$	0.445	0.415
10	$12-6$	0.428	0.403

表 5: Movie Reviews における I_{aw} および AUPRC (acc=0.935, $k_d = 0.2$)

Books (acc=0.900)		DVD (acc=0.878)		Livedoor (acc=0.968)	
層-ヘッド	I_{aw}	層-ヘッド	I_{aw}	層-ヘッド	I_{aw}
α^{merge}	**0.320**	α^{merge}	**0.400**	α^{merge}	**0.457**
$11-11^{NEG}$	0.299	$11-11^{POS}$	0.338	$10-7^{sp,m,d,i}$	0.388
$12-10^{POS}$	0.273	$11-2$	0.303	$12-8^{h}$	0.394
$12-9$	0.266	$10-8$	0.287	$11-8$	0.371
$10-9$	0.243	$12-9$	0.269	$11-1$	0.36
$11-2$	0.213	$10-6^{NEG}$	0.237	$10-4$	0.359
$11-1$	0.187	$10-9$	0.231	$12-avg$	0.358
$12-4$	0.177	$11-9$	0.229	$12-3^{n,t}$	0.351
$11-12$	0.176	$10-1$	0.212	$11-9^{k}_{r=12}$	0.331
$12-avg_{r=20}$	0.116	$12-avg_{r=16}$	0.15	$10-9^{sm}_{r=23}$	0.249

表 6: 根拠のアノテーションが無いデータセットにおける I_{aw} ($k_d = 0.2$)

2 つのアテンションを可視化した結果をランダムに上下に並べてアノテータ [6] に提示し，アノテータは，解釈性について，1 つめ，2 つめ，どちらも高い，どちらも低い，の 4 つの選択肢の中から 1 つを選択した．アテンションの重みは，上位 20%のトークンの背景色に色付けをして可視化した．集計は，どちらも高い場合は両方 1，どちらも低い場合は両方 0 として，個々のサンプルについてのアノテータ 3 名の結果の合計が 2 以上になった場合にカウントする方法をとった．結果を表 7 に示す．全体で 12-avg の 46%に対し，α^{merge} は 92%となり，I_{aw} の有効性が確認された．カテゴリごとに見ると，dokujo-tsushin のみ 12-avg のほうが良い結果となったが，その他のカテゴリについては同数もしくは α^{merge} のほうが良い結果であった．

5 おわりに

本研究では，BERT の 144 のアテンションヘッドについて，文書分類の予測の解釈性を定量的･定性的の両面から分析した．定量的な評価から，相対的に解釈性の高い重みを持つアテンションヘッドが発見された．さらに，根拠個所の正解データが付与されていないデータセットを用いた定性的評価から，提案手法により，アテンションの可視化による予測根拠提示の解釈性を改善できることが示された．これらの結果は，解釈可能な重み自体を出力するモデルの構築に寄与すると考える．

[6] 著者ら 3 名にて実施した．

カテゴリ	12-avg	α^{merge}	データ件数
ALL	46%	92%	100
dokujo-tsushin	57%	29%	7
it-life-hack	31%	85%	13
kaden-channel	47%	100%	15
livedoor-homme	100%	100%	5
movie-enter	57%	95%	21
peachy	0%	100%	5
smax	0%	100%	11
sports-watch	25%	100%	12
topic-news	100%	100%	11

表 7: 定性評価結果．データ件数中の解釈性があった件数の比率

参考文献

[1] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *ACL*, 2019.

[3] Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In *ACL*, pp. 2931–2951, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS*, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

[6] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention Interpretability Across NLP Tasks. *arXiv*, 1909.11218, 2019.

[7] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *ACL*, pp. 5797–5808, 2019.

[8] Paul Michel, Omer Levy, and Graham Neubig. Are Sixteen Heads Really Better than One? In *NIPS*, 2019.

[9] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? An Analysis of BERT's Attention. In *BlackBoxNLP*, pp. 276–286, 2019.

[10] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models Equal contribution. *arXiv*, 1911.03429v1, 2019.

[11] Omar F. Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL HLT 2007; Proceedings of the Main Conference*, pp. 260–267, 2007.

[12] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *ACL*, pp. 440–447, 2007.