E-040

BERT による語連想タスクでのシャープレイ値を用いた連想根拠の提示
Presentation of reason for prediction in a word association task by BERT using shapley values

問井 拓海† 相馬 佑哉† 堀内 靖雄† 黒岩 眞吾†
Takumi Toi Yuya Soma Yasuo Horiuchi Shingo Kuroiwa

1. はじめに

近年，説明可能な AI（XAI）と呼ばれる分野において，深層学習モデルの予測の根拠を示す試みが行われている．本稿では，日本語 Wikipedia で学習した BERT[1,2]を用いて 5 つの刺激語から 1 語の正解語を予測するタスク（以下，語連想タスクと呼ぶ）において，連想の根拠として正解語の予測に貢献した刺激語を順序付けする手法を検討した．順序付けの手法としては，attention 機構（以下，attention と表記）の値を用いる方法に加え，協力ゲーム理論のシャープレイ値[3]を用いる方法を提案し比較を行った．

2. BERT を用いた語連想における根拠提示

本稿では，語連想タスクにおいて BERT の Masked Language Model を用いて MASK に予測語が出現する文（以下，予測文と表記）と人が想定した正解語を使用して，正解語に対する連想根拠の提示を行う．具体的には，

＜*刺激語 1*＞、＜*刺激語 2*＞、＜*刺激語 3*＞、＜*刺激語 4*＞、
＜*刺激語 5*＞ *から連想される国は「MASK」です。*

等の文を与え，MASK に予測語が出力されるようにした（文中，*国* の部分は正解語のカテゴリーを指定する語となる）．その上で正解語の予測に貢献した刺激語を順序付けする．順序付けでは各刺激語が正解語を予測した「貢献度」として以下の節で示す attention の値，および，シャープレイ値を用いた手法を検討する．

語連想タスクとしては文献[4]中の『まとめる語想起』132 題のうち，20 代男子大学生 3 人全員の解答が正解の連想語（以下，正解語）と一致し，かつ東北大 BERT[4]で正解語が 1 つの MASK トークンで出力可能な 87 題を使用した．表 1 に実験で使用した予測文を示す．表中の〇は刺激語を表す．なお，文献[5]で MASK に鍵括弧「」を付与することで連想結果が改善されたことから本稿でもそれを踏襲した．

表 1 実験で使用した予測文

〇、〇、〇、〇、〇は「MASK」の仲間です。
〇、〇、〇、〇、〇からできているものは「MASK」です。
〇、〇、〇、〇、〇から連想する色は「MASK」です。
〇、〇、〇、〇、〇から連想する都道府県は「MASK」です。
〇、〇、〇、〇、〇から連想する国は「MASK」です。
〇、〇、〇、〇、〇から連想するスポーツは「MASK」です。
〇、〇、〇、〇、〇から連想する季節は「MASK」です。
〇、〇、〇、〇、〇から連想する場所は「MASK」です。
家の中で〇、〇、〇、〇、〇がある場所は「MASK」です。
〇、〇、〇、〇、〇から連想する行事は「MASK」です。
〇、〇、〇、〇、〇を使ってすることは「MASK」です。
〇、〇、〇、〇、〇は「MASK」のときに持っていきます。
〇、〇、〇、〇、〇から連想するメニューは「MASK」です。

2.1 attention の値を用いて順序付けする方法

本節では，深層学習モデルの予測の根拠を示す方法として広く行われている attention の値を用いて順序付けする方法を説明する．本稿では BERT の最終層の 12 個からなる Multi-Head Self-Attention の出力値を合計し，その中で MASK トークンを query とした際の各刺激語の attention の値を，正解語の予測に貢献した「貢献度」とする．attention の値は文中の刺激語の位置によって変化するため，5 つの刺激語の並び順の総数（5！通り）で attention の値を計算し，平均を取ることで位置による影響を除いた．また，サブワード分割された刺激語，複数の単語に分割された刺激語については，サブワードもしくは複数に分割された刺激語の attention の値を合計して，その刺激語の「貢献度」とした．

2.2 シャープレイ値を用いて順序付けする方法

本節では提案手法である，協力ゲーム理論のシャープレイ値の概念を用いて正解語の予測に貢献した刺激語を順序付けする手法を説明する．シャープレイ値とは，複数人が協力して得た報酬がある場合に，各個人が報酬に対してどの程度貢献したかを表す「貢献度」を示す値である．シャープレイ値での「貢献度」は，各個人が参加したときとしなかったときで，報酬がどの程度増えるか，を基に算出される．本稿における報酬とは，MASK トークンにおける正解語のスコア（*Softmax* 関数を適用する前の値，logits）であり，各個人とは，予測文中の刺激語である．

「貢献度」の計算では，まず 5 つの刺激語の並び順の総数（5！通り）の各々に対し，刺激語を後方から順に削除した 1〜5 語の刺激語で構成される 5 通りの文を考え，各文が予測結果として正解語を出力するスコアを計算しておく．その上で，＜刺激語 1＞ が含まれていない文と，その文に＜刺激語 1＞ を刺激語並びの最後尾に追加した文のスコアの差分を求める．また，＜刺激語 1＞ が含まれていない文が，同一の刺激語の組み合わせで構成されている点も考慮して，＜刺激語 1＞ の追加によるスコアをそれら全ての組み合わせのスコアの差分の平均として算出する．ただし，本稿では予測をしたときに正解語の順位が 101 位以下になる場合は，正解語のスコアを 0 とした．以上の処理により求めた平均が本稿で提案する<刺激語 1> のシャープレイ値であり，<刺激語 1> 単独の「貢献度」になる．同様の手順で <刺激語 2> 〜 <刺激語 5> でも計算を行い，シャープレイ値（「貢献度」）の大きい順に刺激語を順序付けする．

3. attention の値とシャープレイ値の比較実験

3.1 評価指標

順序付けの妥当性を客観的に判断する評価指標として以下の方針で求めた平均逆順位（*MRR*）を用いる．

順序付けした刺激語のうち，下位の刺激語は正解語の予測に貢献している度合いは小さいと考えられる．そのような刺激語を除いて予測を行っても正解語の順位が下がらな

† 千葉大学 Chiba University

い場合，「貢献度」の大小を正しく判別できていることになる．このことを確認するため，順序付けした刺激語の上位 1 語，2 語，…，5 語を用いて各々予測を行い，正解語の順位を基に各々で *MRR* を計算し評価指標とした．

MRR とは，ランク付けされた結果を返すモデルの評価に用いられる指標であり，式(1)にその定義を示す．

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \quad (1)$$

ここで，N は予測文の数，$rank_i$ は各予測文において正解語が初めて出現した順位である．*MRR* は 0 から 1 の値を取り，1 に近い方法の方が，BERT が正解語を上位に出力できていることを示している．つまり，上位の刺激語だけで予測を行なった場合でも *MRR* が大きければ，予測に貢献している刺激語を正しく順序付けできていると判断できる．なお $rank_i$ の最大値は本タスクの実用性を考慮した 5 と一般的に用いられる 100 とし，それより大きい場合は $1/rank_i$ を 0 として *MRR* を求めた．

3.2 実験結果

表 2 に各々の方法（attention，シャープレイと表記）における，$rank_i$ の最大値を 5 および 100 とした場合の，順序付けした刺激語の上位 1〜5 語を用いて予測した結果の *MRR* を示す．全ての条件で，シャープレイ値の *MRR* が attention の値による *MRR* を上回っていることがわかる．

表 2 $rank_i$ の最大値 5, 100 における *MRR*

$rank_i$ 最大値	刺激語数	5	4	3	2	1
5	attention	0.420	0.417	0.411	0.338	0.279
	シャープレイ	**0.434**	**0.471**	**0.484**	**0.401**	**0.331**
100	attention	0.440	0.435	0.427	0.363	0.295
	シャープレイ	**0.452**	**0.494**	**0.501**	**0.424**	**0.352**

3.3 考察

シャープレイ値による方法で *MRR* が高くなった要因は，正解語の順位を正解語のスコアがどの程度高くなるかで決定した点にあると考えられる．すなわち，刺激語を削る際，正解語のスコアの上昇に関与しない刺激語を選んだことによるものと考えられる．

またシャープレイ値による方法では，刺激語の上位 3，4 語を用いた場合に，5 語（刺激語を削除していない予測文）での *MRR* を上回った．これは本稿で用いた予測文において，シャープレイ値による「貢献度」が負になる刺激語が存在し，正解語の予測に負の貢献をしているためである．具体例で説明するために，表 3 に，正解語が楽器の場合の各方法での刺激語の順序を，表 4 に正解語が楽器の場合の刺激語上位 1〜5 語で予測を行ったときの正解語の順位を示す．表 3 に示したように，シャープレイ値では刺激語の順序は <バイオリン>，<琴>，<太鼓>，<フルート>，<ピアノ> である．このとき，5 語の予測による正解語の順位が 22 位であるのに対し，<ピアノ> を除くことで 7 位，さらに <フルート> を除くことで 2 位となった（表４）．このことから <フルート>，<ピアノ> は，正解語の予測に対して負の貢献をしていたと判断できる．上位 3 語で *MRR* が最も高くなった要因は，この例のように負の貢献をする刺激語が 2 つある場合が多いからであると考えられる．しかし実際に求めた「貢献度」は，<ピアノ> は負に，<フルート> は正になっており，<フルート> は正解語の予測に正の貢献をすることが期待されていた．この要因として刺激語の位置の影響（<フルート> は平均的には正の貢献をするが表 3 の順序では負の貢献となる）があると考え，表 3 の順序で <フルート> のみを削除して 4 語で予測を行い、5 語で予測した場合と正解語のスコアを比較した．その結果，＜フルート> を除くことで正解語のスコアは低下しており，＜フルート＞ は正の貢献をしていると考えられる．このことから位置の影響ではないと判断できる．

その他に考えられる要因は，<フルート> と組み合わせると，正解語の予測に負の貢献をしてしまう刺激語が存在する可能性である．組み合わせの影響は，表 4 において上位 2 語を用いたときにシャープレイ値を用いた方法の順位を attention よりも低くしている点にも現れている．すなわち，シャープレイ値が単独の刺激語の貢献度を表している一方で，本稿では複数の刺激語が組み合わさった貢献度をその和として考えていることが原因と考えられる．例えば <太鼓> と <琴> の 2 語の組み合わせに注目してスコアの差分は計算していない．こうした組み合わせの影響を考慮した貢献度の計算方法の構築は，今後の課題である．

表 3 正解語が「楽器」のときの刺激語の順序

attention	太鼓	琴	バイオリン	フルート	ピアノ
シャープレイ	バイオリン	琴	太鼓	フルート	ピアノ

表 4 表 3 の順序で刺激語を提示した予測文での正解語「楽器」の順位

刺激語数	5	4	3	2	1
attention	18	9	**4**	14	ランク外
シャープレイ	22	7	**2**	27	42

4. おわりに

本稿では BERT による語連想タスクにおいて，連想の根拠としての刺激語を順序付けし各語の貢献度を提示する手法の検討を行った．順序付けには attention の値を用いる方法と，協力ゲーム理論のシャープレイ値を用いる方法の 2 つを用いた．貢献度順に提示した 1〜5 語の刺激語で連想を行い，*MRR* を計算した結果，シャープレイ値による方法の *MRR* が高く，同手法が正解語の予測に貢献した刺激語の順序付け方法として妥当であることがわかった．今後は語の組み合わせによる貢献度も考慮した順序付け手法を検討する．

謝辞

本研究を進めるに当たり，乾・鈴木研究室の訓練済み日本語 BERT モデル[2]をお借りしました．モデルを公開してくださったことに厚く御礼を申し上げ，感謝の意を表します．本研究は JSPS 科研費 JP20K11860, JP21K02052 の助成を受けたものです．

参考文献

[1] Jacob Devlin, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019.
[2] 東北大学乾・鈴木研究室．“Pretrained Japanese BERT models,” https://github.com/cl-tohoku/bert-japanese，引用日: 2022/06/24.
[3] L. S. Shapley. “A Value for n-Person Games,” pp. 307–318. Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, 1953.
[4] 鈴木勉 宇野園子監修．CD 版そのまま使える 失語症教材 2．エスコアール，2022 出版予定.
[5] 相馬佑哉，他, “人間と BERT の語から語の連想の比較,” 言語処理学会第 27 回年次大会，2021.