Vol.2021-NL-249 No.1
2021/7/27

深層学習による文書の話題分類の判断根拠の提示に関する一考察

為栗敦生[1]　中村鴻介[2]　高橋良颯[3]　山口実靖[4]

概要：深層学習は文書分類等の自然言語処理にて活用され，Self-Attention などが大きな成果をあげている．一方で深層学習による分類は，分類精度は高いがその判断根拠を人間が理解することが困難であるとの指摘がされている．本稿では，テーマが定められたニュース記事群のテーマによる分類のタスクに着目し，深層学習による分類の判断根拠の提示手法について考察する．具体的には，LSTM Attention により記事分類を行い，高い精度で分類をできることを示す．そして，Attention 値や既存の判断根拠提示手法 Smooth-grad に着目し，自然言語記事分類の判断根拠提示手法について考察する．また性能評価により，これらに着目することにより判断根拠を提示できることを示す．

キーワード：機械学習，深層学習，記事の分類，self-attention，Smooth-grad

A Study on Presenting Decision Rationale for Topic Classification of Documents by Deep Learning

ATSUKI TAMEKURI[†1]　KOSUKE NAKAMURA[†2]
YOSHIHAYA TAKAHASHI[†3]　SANEYASU YAMAGUCHI[†4]

Abstract: Deep learning has been used in natural language processing (NLP) such as document classification. In particular, Self-Attention has achieved significant results in NLP. However, it has been pointed out that although deep learning highly accurately classifies documents, it is difficult to interpret the basis of the decision. In this paper, we focus on the task of classifying news documents by their theme. We then propose methods for presenting the interpretability for classification decisions by deep learning. We first classify documents with LSTM Attention and show that this can classify documents with high accuracy. We second propose five methods for providing the basis of the decision by focusing on various values, e.g. Attention. Finally, we evaluate the methods and show that these methods can present interpretability.

Keywords: Mechanical Learning，Deep Learning，Article classification，self-attention，Smooth-grad

1. はじめに

近年，深層学習の利用により自然言語処理などにおいて複雑なタスクの処理がより高い精度で可能にとなり，深層学習が大きな注目を集めている．深層学習ではタスクに応じて，多層に組み合されたニューロンにより深層ニューラルネットワーク(DNN, Deep Neural Network)，畳み込みニューラルネットワーク(CNN, Convolutional Neural Network)，再帰ニューラルネットワーク(RNN, Recurrent Neural Network)，RNN を拡張した LSTM(Long Short-Term Memory)などを構成することが可能である．深層学習は，高度な分類や推論が必要とされる自然言語処理や画像認識などの分野において，既存の機械学習手法と比べて高精度な分類や推論を可能にしている．しかし，文献[1][2]などにおいて，深層学習はブラックボックスであり推論結果に対する解釈性や説明性が無いため，信頼することが困難であるという欠点が指摘されている．人間の社会生活において判断の説明や解釈が求められる場面は数多く存在する．例えば，裁判における判決に対しての判断根拠の提示や，経営者から株主への経営戦略の判断理由の説明や，政治家の政治判断の有権者への説明などが考えられる．これらのように判断結果に対する解釈性や説明性の付与は重要であると言える．

本稿では，テーマが定まった 2 種類のニュース記事の分類タスクに着目し，self-attention による記事の分類モデルを構築することで記事の分類を行う．そして，その分類の判断根拠の提示手法についての考察を行う．具体的には，判断根拠の提示手法を 5 つ提案し，それらが提示した判断根拠の評価を行う．

2. 関連研究

2.1 self-attention による自然言語処理

self-attention[3][4]は Bidirectional LSTM に Attention を組み合わせることで構成されるニューラルネットワークモデルである．Bidirectional LSTM は，1 つの入力に対して順方向と逆方向の 2 方向から計算を行い，計算によって得られた順方向 LSTM の出力値と逆方向 LSTM の出力値の 2 つを連結することで，Bidirectional LSTM の出力としている．これにより，Bidirectional LSTM は学習および分類を行う際に，順方向 LSTM から前の情報を，逆方向 LSTM から後の情報を得ることができ，前後の文脈から単語に関する情報を加えることが可能である．

2.2 機械学習における判断根拠の提示

深層学習の判断根拠を示す手法として，ノイズの追加による顕著性マップの作成によって注目する次元を推定する

SmoothGrad[5]がある．SmoothGrad は，深層学習(主に CNN)の入力値にガウシアンノイズを加えることで入力次元毎に入力値の変化に対する出力値の変化(勾配値)を計算し，判断根拠となる勾配値が大きい部分を入力された情報から抽出する手法である．

Ribeiro らは，機械学習において分類の決定を解釈することが重要であると主張している[2]．同文献では，機械学習モデルの多くがブラックボックスであることを指摘し，解釈性や説明性の付与が重要であると主張している．また著者らは，写真内の動物がシベリアンハスキーか狼かの判断をする学習モデルの例を示し，その判断根拠が「画像の背景が雪であるか否か」であることを述べ，この学習モデルは“Bad model”であると主張している．そして，ブラックボックスである機械学習モデルの判断根拠の理解は，その信頼性を評価する際に重要であると主張している．

中村らは，深層学習などの機械学習による自然言語処理における解釈性の付与手法[6][7]を提案している．文献[6]では，SVM および DNN を用いてレビュー情報の高評価であるか低評価であるかの分類を行い，その分類における判断根拠の提示手法として SVM の重みベクトルの絶対値や，SmoothGrad を用いる手法を提案している．同文献では，分類精度が高い場合であっても人間の主観的に適切とされる判断は行われておらず，“Bad Model”に近い根拠で判断されている可能性もあると述べている．文献[7]では，self-attention による 2 つのニュース記事の分類モデルの構築を行い，記事分類に判断根拠を付与する手法として SmoothGrad を自然言語処理に拡張した手法や，Attention 値を用いる手法を提案している．同文献では，SmoothGrad を拡張した手法よりも Attention 値を用いる手法の方がより重要である判断根拠語を抽出できていると述べている．

3. self-attention によるニュース記事の分類

本稿では，深層学習によりニュース記事の分類を行い，この分類を判断根拠提示の対象として考察を行う．本章にて，このニュース記事分類について説明する．

3.1 分類対象と分類方法

まず分類で用いたニュース記事について述べる．分類対象には livedoor ニュースコーパス文献の 9 ジャンルの中から「家電チャンネル」と「エスマックス」の 2 種類を選択し，これらの記事に対して Self-attention による機械学習モデルを用いて分類し，その記事が「家電チャンネル」のものであるか「エスマックス」のものであるかの推定を行った．9 ジャンルの中から「家電チャンネル」と「エスマックス」を選択した理由は次節で述べる．それぞれの記事は，家電とモバイルガジェットの話題を取り扱ったニュース記事である．学習に用いた記事数は家電チャンネル 864 記事と，エスマックス 870 記事であり，ニュース記事の形態素解析には MeCab 0.996 を使用し，MeCab 辞書には NEologd を使用した．単語のベクトル表現には fastText を用いており，ウィキペディア日本語コーパスの事前学習モデルを fastText のモデルとして用いた．また，分類対象の記事に対して品詞や記号などの除外は行っていない．分類に用いた Self-attention の機械学習モデルのハイパーパラメータ等は，フレームワーク Pytorch 1.3.1, Bidirectional LSTM 出力次元数 512, 単語ベクトル次元数 300, 最適化関数 Adam，学習率 0.001, バッチサイズ 128, 損失関数 CrossEntropy とした．機械学習モデルの作成に用いた訓練データとテストデータは 80%:20% の比率とし，学習は訓練データの損失関数が収束するまで行った．

3.2 分類精度

self-attention を用いた学習を行い 2 種類の記事を分類した精度(accuracy)は 99.1%であった．テストデータ数は「家電チャンネル」が 174，「エスマックス」が 173 であり，「家電チャンネル」を「エスマックス」と誤って分類したものが 1 つ，逆方向に誤って分類したものは 2 つであった．

livedoor ニュースコーパスは 9 ジャンルあるため生地分類の 2 ジャンルの選択の組み合わせは ${}_9C_2 = 36$通りある．これら 36 通りの組み合わせの中で，「家電チャンネル」と「エスマックス」の精度が最も高い結果となった．本研究ではまず，深層学習が正しい判断を行えるが解釈性がないことに着目して考察を行う．よって，本稿では最も精度が高かった本ジャンルの分類に着目して考察を行う．精度が低いジャンルの分類に関する考察は，分類を誤る原因の特定や，精度向上に関する考察として有用であると予測される．

4. 判断根拠提示手法

本章にて，分類に用いる判断根拠の提示手法として，NLG 絶対値手法，Attention 値手法，WD 値手法，NLG*WD 値手法，Att*WD 値手法の 5 つの手法を提案する．

4.1 NLG 絶対値

NLG 絶対値手法は，SmoothGrad[5]をナイーブに Self-attention による自然言語処理に適用した手法である．ベクトル表現に変換された入力文に対してノイズを付加することで，付加したノイズの値による出力値の変化を調査し，各次元における入力語データ(ベクトルのある要素)に対する出力値である勾配値を求める．勾配の値は入力ベクトルの要素毎にあるため，勾配はベクトルとなり，勾配ベクトルの次元数と入力データのベクトルの次元数は等しくなる．本稿では，このベクトルの勾配値の絶対値を NLG 絶対値と呼び，本手法はこの NLG 絶対値が大きい語を，分類において重要である判断根拠語とする．NLG 絶対値が大きい語は分類に大きな影響を与える語であると考えることが出来るが，その語が判断結果を true の方向に大きく変える影響の大きい語であるか，false の方向に大きく変える影響の大きい語であるのかは考慮されていないと言える．

4.2 Attention 値

Attention 値手法は，ベクトル表現変換された入力文に対して Attention の計算を行い，各単語に出力された Attention 値を用いる手法である．本手法は出力された Attention 値の絶対値が大きい語を，分類において重要である判断根拠語とする．

4.3 WD 値

WD 値手法は，ベクトル表現に変換された入力文内の 1 単語を self-attention による機械学習モデルに入力し，その出力値を用いる手法である．入力文書内の語を 1 つずつ抽出し，その 1 語のみで構成される文を作成し，その 1 語文を当該モデルに入力し，その出力値を得る．この出力値を本稿では WD 値と呼ぶ．WD 値は各語のそのモデルにおける評価を表現する値と考えることが出来るが，文脈は考慮されていないと言える．本手法は，この WD 値が大きい語を分類において重要である判断根拠語とする．対象の語が false の分類結果を象徴する語である場合は，その語の WD 値は負の値となる．

4.4 NLG*WD 値

NLG*WD 値手法は，各単語の NLG 絶対値と，WD 値の符号(+1 または-1)を求め，NLG 絶対値(正の実数)と WD 値符号(+1 または-1)の積を判断根拠の語とする手法である．NLG 絶対値がその語の判断根拠としての大きさを，WD 値符号がその判断根拠としての方向を表現していると考えることが出来る．

4.5 Att*WD 値

Att*WD 値手法は，NLG*WD 値と同様に各単語の Attention 値と，WD 値を求め，Attention 値(正の実数)と WD 値符号(+1 または-1)の積を判断根拠としての重要語とする手法である．Attention 値がその語の判断根拠としての大きさを，WD 値符号がその判断根拠としての方向を表現していると考えることが出来る．

5. 性能評価

5.1 評価方法

提案した判断根拠提示手法のニュース記事分類タスクにおける性能を，4 種類の方法(評価方法 1-1, 1-2, 2-1, 2-2)を用いて評価した．「家電チャンネル」と「エスマックス」のそれぞれから無作為に 5 記事ずつを選択し，それらを検証対象とした．これら検証対象の記事の分類を 3 章の self-attention による機械学習モデルを用いて行い，その出力値を得た．分類結果は，10 記事とも正しい分類であった．次に，これら 10 件の記事の分類に対して 4 章の各手法により判断根拠の提示を行った．

評価方法 1-1 では，各手法が判断根拠として重要と示した語を，各手法が提示する重要度が高い語から順に記事から削除していき，単語がされた記事を当該モデルを用いて分類していく．当該モデルが判断を誤るまで単語の削除を繰り返していき，何語削除したら判断を誤るかにより各手法の性能を評価する．少ない削除で判断を誤った場合は，それらの少ない語が重要な判断根拠であったと考えることができ，その手法はより優れた判断根拠を提示することができたと考えることができる．ただし，単語削除を繰り返すと，元の記事と大きく異なる文書となってしまい，その元の記事と大きく異なる文書における定量的評価の重要性は低いとも考える事ができる．

評価手法 1-2 では，手法 1-1 の「元の記事と大きく異なる文書」という状況における評価を避けるため，元も記事から各手法が判断根拠として重要であると示した単語のうちの上位 10 件までを削除し，単語の削除により self-attention による機械学習モデルの分類結果の出力値がどの程度変化するか(出力値が判断を誤る方向にどの程度変化するか)により評価する．

評価方法 2-1 と 2-2 では，各手法により記事内の各単語に重要性の値を付与し，その重要性の順位表を作成する．当然，重要な語を順位表の上位に位置させる手法が優れた判断根拠提示手法であると言える．次に，この各手法が提示した順位表と，以下に述べる正解と仮定した順位表とを比較し，両表が近いほど優れた提示手法であると仮定して評価する．

正解と仮定した順位表は，以下の手順で作成する．まず，検証対象の記事からある 1 語を削除した文書を作成する．そして，その文書を self-attention 機械学習モデルにより分類し，分類結果の出力値を得る．単語削除前の記事の分類結果出力値と，削除した記事の分類結果出力値を比較する．本稿ではこの変化量をその単語の削除による評価変動値と呼ぶ．文書内の全単語について評価変動値を求め，評価変動値が分類を誤る方向に大きい語から順に，分類の判断根拠における重要な語として順位表を作成する．この順位表を，正解と仮定した順位表とする．

評価方法 2-1 では，各手法が提示した順位表と，正解と仮定した順位表の順位の値の相関係数を求め，相関係数が高いほど優れて判断根拠の提示手法とする．評価方法 2-2 では，評価変動値の高い上位 30 語(すなわち正解仮定した順位表の上位 30 語)に対する各提示手法の順位表の nDCG 値を計算することで性能評価を行う．nDCG 値が高いほど，優れた根拠提示手法と考える．

5.2 性能評価結果

評価方法 1-1 の，「家電チャンネル」と「エスマックス」のそれぞれ 5 記事，検証対象全体の 10 記事における各提示手法が分類を誤るまでに削除した単語数の割合の平均を図 1 に示す．図の縦軸が，検証対象の記事の全単語数を 100%とした場合における分類を誤るまでに削除した単語数の割合を示示している．図 1 より，「エスマックス」では WD 値手法が 16.2%で，「家電チャンネル」では Att*WD 値手法が 11.0%で，最も少ない削除割合で分類を誤る提示手法と

なっていることが分かる．検証対象全体の 10 記事の平均では WD 値手法が 14.0%となり最も少ない削除割合で分類を誤る提示手法となっていることが分かる．また，NLG 絶対値手法が全 10 記事の平均で最も高い(悪い)結果であり，検証対象の記事から 80%以上の単語を削除しなければ分類を誤らなかったことが分かる．以上のことから，分類を誤るまで単語を削除する検証方法においていは，WD 値手法が他の提示手法よりも判断根拠として重要性の高い単語を抽出したと言える．また，検証対象の記事群の中から「家電チャンネル」と「エスマックス」からそれぞれ 1 記事を選択し，分類を誤るまでの評価値の推移を図 2, 図 3 に示した．図 2,

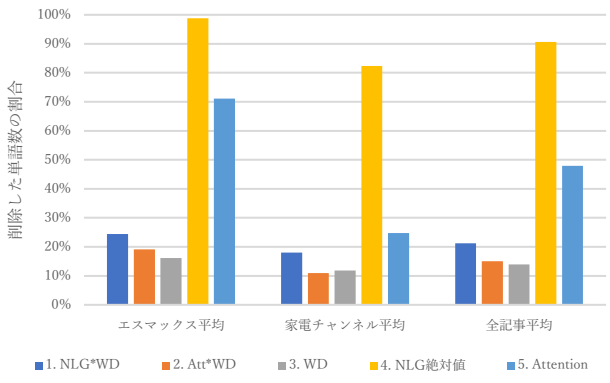

図 1　各提示手法における削除単語数割合

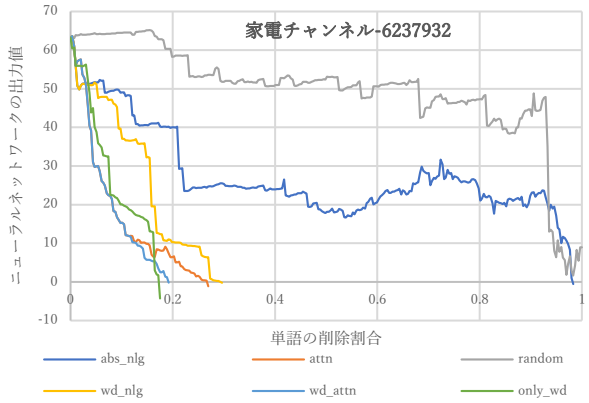

図 2　家電チャンネルの評価値推移

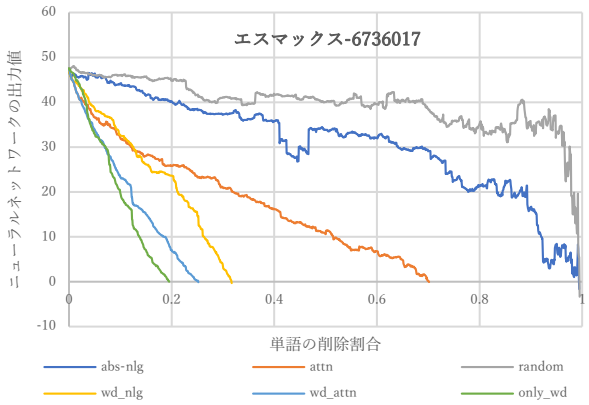

図 3　エスマックスの評価値推移

図 3 の縦軸は評価値であるニューラルネットワークの出力値であり，横軸が削除した単語数の割合である．比較のために，ランダムな順で削除した場合の評価値の変化も示す．図より，図 1 において性能が優れる NLG*WD 手法，Att*WD 手法，WD 値手法は少ない対語削除で評価値が急速に減少していることが分かる．

評価方法 1-2 の評価結果を図 4 に示す．図の各線は各手法の 10 記事の評価結果の平均を表しており，横軸は各提示手法の評価値の高い単語の削除数(1 語から 10 語)を示している．

図 4 より，削除単語数が 1 語から 10 語の全てにおいて，Att*WD 値手法が最も評価値の変化の割合が大きいことが分かり，最も優れた手法であると考えることができる．また，評価方法 1-1 で最も良い結果であった WD 値手法は，評価値の高い 10 語が削除対象の場合では全提示手法内で 4 番目に良い手法であることが分かる．前節で述べた様に，元の記事から多数の単語を削除した文書の定量評価結果は重要でないと考えることができ，この考えに基づくと，判断根拠としての重要性が高い単語を抽出する提示手法としては，評価方法 1-1 で優れる WD 値手法より，評価方法 1-2 で優れる Att*WD 値手法が優れた提示手法であると考えることができる．また，検証対象の記事群の中から「家電チャンネル」と「エスマックス」からそれぞれ 1 記事を選択し，評価値の高い 10 語を削除時の評価値の推移を図 5, 図 6 に示す．

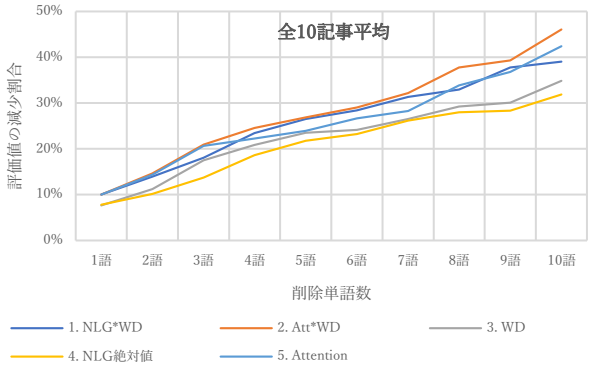

図 4　評価値の高い 10 語削除時の評価値の推移

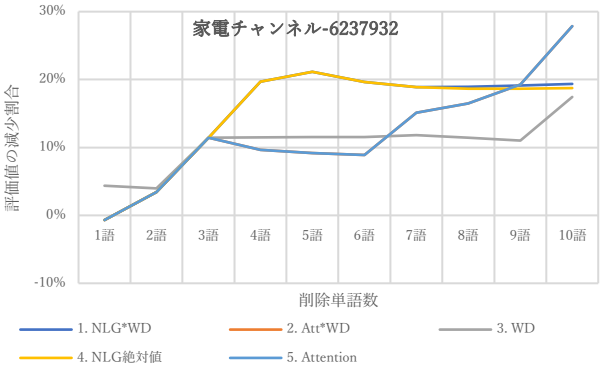

図 5　家電チャンネルの 10 語削除時の評価値の推移

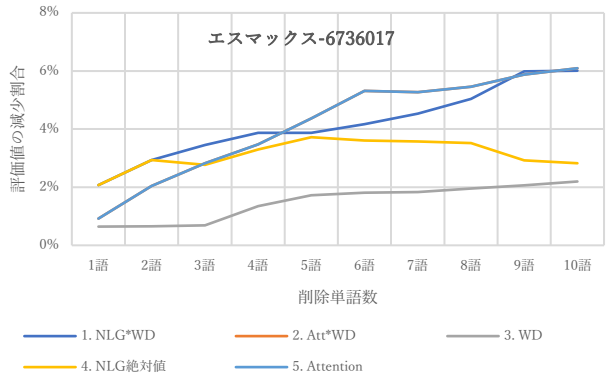

図 6　エスマックスの 10 語削除時の評価値の推移

評価方法 2-1 の評価結果(相関係数)を図 7 に示す．本稿の記事分類では，深層学習がエスマックスと分類した場合は正の値を，家電チャンネルとして分類した場合は負の値を出力することとなっている．よって，エスマックス記事の相関係数は大きな値であるほど，家電チャンネル記事の相関係数は小さな値であるほど(絶対値の大きな負の値であるほど)優れた判断根拠提示になっていると言える．また，エスマックスの相関係数と，家電チャンネルの相関係数を-1 倍したものの平均を，図の全記事として示している．

図 7 にて，「エスマックス」では検証対象の 5 記事の相関係数の平均が，NLG*WD 値手法にて 0.65，Att*WD 値手法にて 0.63，WD 値手法にて 0.71，NLG 絶対値手法にて 0.03，Attention 値手法にて 0.07 となっており，WD 値手法が最も強い正の相関があり，最も優れていることが分かる．「家電チャンネル」では提示手法による向きの付与が「エスマックス」とは逆になっているため，負の相関が表れていることが分かる．検証対象の 5 記事の相関係数の平均では，NLG*WD 値手法が相関係数-0.54，Att*WD 値手法が相関係数-0.47，WD 値手法が相関係数-0.51，NLG 絶対値手法が相関係数-0.06，Attention 値手法が相関係数-0.11 となっており，NLG*WD 値手法が最も強い負の相関があり，最も優れていることが分かる．また，「家電チャンネル」の相関係数に対して-1 をかけて全 10 記事での相関係数の平均を計算した全記事平均では，NLG*WD 値手法が相関係数 0.60，Att*WD 値手法が相関係数 0.55，WD 値手法が相関係数 0.61，NLG 絶対値手法が相関係数 0.04，Attention 値手法が相関係数 0.09 となっており，WD 値手法が最も良い相関があることが分かる．また，検証対象の記事群の「家電チャンネル」と「エスマックス」の中からそれぞれ 1 記事ずつ選び，単語 1 語を削った状態の評価値である評価変動値の順位表と，それ対する各提示手法の単語の評価値の順位表の相関係数を図 8 に示す．

評価方法 2-2 の評価結果を図 9 に示す．図の縦軸は，「家電チャンネル」と「エスマックス」のそれぞれ 5 記事と，検証対象全体の 10 記事における評価変動値の高い上位 30 語に対する各提示手法の単語の評価値の順位表の正解と仮定

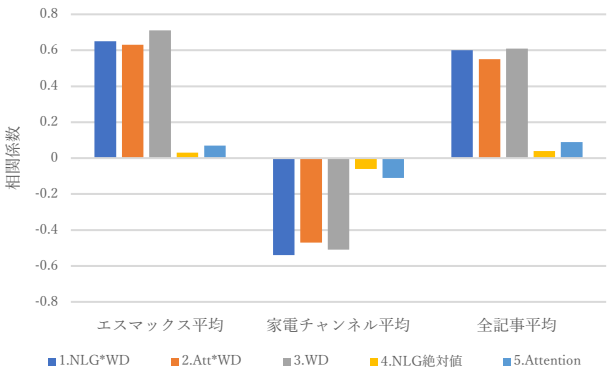

図 7　評価変動値に対する各提示手法の相関係数

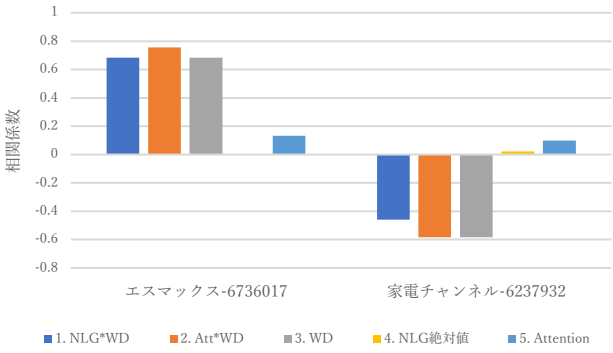

図 8　家電チャンネルとエスマックスの相関係数

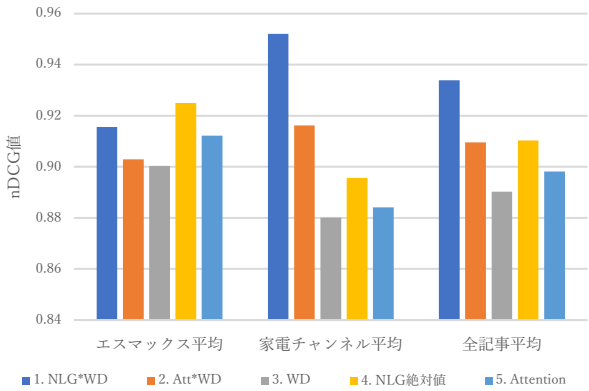

図 9　評価変動値に対する各提示手法の nDCG 値

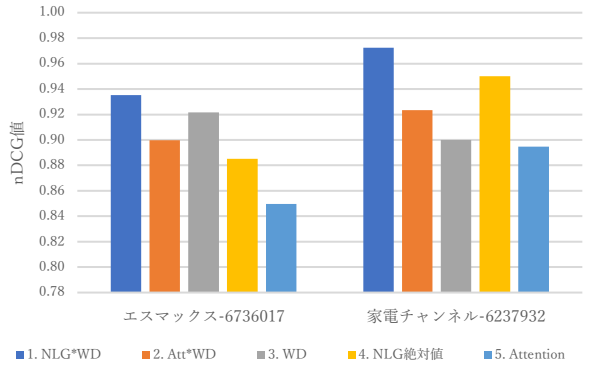

図 10　家電チャンネルとエスマックスの nDCG 値

した順位表に対する nDCG 値の平均である．図 9 より，「エスマックス」では NLG 絶対値手法が最も高い nDCG 値であり NLG*WD 値手法が 2 番目に高い nDCG 値であり，「家電チャンネル」では NLG*WD 値手法が最も高い nDCG 値であることが分かる．また，検証対象全体の 10 記事の平均では NLG*WD 値手法が最も高い nDCG 値であることが分かる．また，検証対象の記事群の「家電チャンネル」と「エスマックス」の中からそれぞれ 1 記事ずつを選び，それらの記事の評価変動値の高い上位 30 語に対する各提示手法の単語の評価値の順位表の nDCG 値を図 10 に示す．

以上の 4 種類の評価方法を用いて性能評価を行った結果，評価方法 1-1 と評価方法 1-2 の評価値の高い単語を選択し抽出するには Att*WD 値，評価値が低い単語で構成された記事の分類では WD 値が最も良い結果となった．また，評価方法 2-1 と評価方法 2-2 の self-attention による機械学習モデルを用いた解釈性の付与の結果に，最も近い結果を出力する提示手法は WD 値手法が，上位 30 語などに限定した場合では NLG*WD 値手法が最も良い結果となった．評価手法 2-1 および 2-2 の様な，特に重要な判断根拠語を抽出するには NLG 値の計算(すなわち SmoothGrad による判断根拠の抽出)が重要であることが分かる．

これらの結果から，WD 値手法が評価方法 1-1 と評価方法 2-1 で最も優れており，概ね多くの評価方法で優れた判断根拠を示すことが期待できる．ただし，評価方法 2-2 などにおいては最も悪い nDCG 値となり，さらなる検証が重要であると考えられる．

6. 考察

WD 値手法は一部の性能評価(評価方法 1-1 や 2-1)にて最も良い提示結果を示した．しかしながら，WD 値は 1 単語の文書を self-attention による機械学習モデルのネットワークに入力して解釈性の付与(判断根拠の重要性の判断)を行っている．よって，機械学習モデルの分類精度が低い場合は WD 値の正確さや優位性も下がると予想される．本稿で用いた機械学習モデルでは分類精度が 99.1%と高かったため，WD 値手法においても良い根拠が抽出できたと考えられる．また，評価値の高い単語の選択と抽出においては Att*WD 値手法が最も良い結果であったため，Attention 値で求めた評価値に高い精度で方向性を付与することができれば，より良い提示手法を実現できると期待できる．

性能評価における評価方法 2-1 では，WD 値に強い相関がみられる理由としては，評価方法 1-1 と同様に分類に用いる機械学習モデルの分類精度が高かったためと考えられる．また，nDCG 値において NLG*WD が良い結果となった理由としては，上位 30 語に単語の評価値が高く方向性が異なっている単語が比較対象に含まれていなかったからだと考えられる．

本稿では，ある語を文書から削除した場合に判断を誤ったり，判断結果の評価値が大きく減少したりした場合にその語を判断根拠の語とみなし，これらの語を正解と仮定した．そして，各提案手法の提示結果と正解の差異を比較し，差異の少ない手法を優れた手法と評価した．しかし，本来はそのモデルが判断根拠としている語と提案手法の提示語の比較を行うことが好ましい．ただし，現在はモデルが判断根拠としている正解を検出する手法が存在しないため，本稿ではこの評価手法を用いた．また，モデルが判断根拠としている正解を抽出する手法が確立されれば，それが本研究の研究目的の達成となる．

また，本稿の評価方法 1-1 から 2-2 は，評価方法が定量的に定義されているため，この評価方法の評価値が高い語を提示する手法が最も高い評価となる．例えば，ランダムに抽出した組み合わせや候補の全ての組み合わせの評価値を求め，評価値が最高のものを提示する手法を実装すれば，その手法が最高の評価値をとなる．ただし，正解と仮定した結果(正解と仮定した順位表など)が，モデルが判断根拠としている正解であることは確認されておらず，そうではない可能性が十分にあると考えられる．よって，今後は本稿で正解と仮定した組み合わせ(正解と仮定した順位表など)を他の手法が提示した判断根拠と比較し，提案手法の提示結果が「本稿で正解と仮定した結果」よりも優れていることの証明などに取り組んで行くべきであると考えられる．

7. おわりに

本稿では深層学習による自然言語処理における判断根拠の提示手法に着目し，5 つの提示手法を提案した．そして，ニュース記事の分類タスクにおける判断根拠の提示の適切さを 4 種類の評価方法を用いて評価した．評価の結果，WD 値を用いる手法が多くの場合により高い適切さで根拠を提示することが可能であると期待できることが分かった．ただし，手法の優劣は評価方法に依存して変化することも確認されており，さらなる検証が重要であると考えられる．

今後は，別の評価手法による評価，さらに多くの記事による評価などを行う予定である．

謝辞

本研究は JSPS 科研費 18K11277, 21K11854 の助成を受けたものである。

本研究は、JST、CREST JPMJCR1503 の支援を受けたものである

参考文献

[1] Grégoire Montavon, Wojciech Samek and Klaus-Robert Müller, Methods for Interpreting and Understanding Deep Neural Networks, Digital Signal Processing Volume 73, Pages 1-15, February 2018.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).

ACM, New York, NY, USA, 1135-1144. DOI: https://doi.org/10.1145/2939672.2939778

[3] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou and Yoshua Bengio, A Structured Self-attentive Sentence Embedding, The International Conference on Learning Representations (ICLR ‘17), 2017.

[4] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, The International Conference on Learning Representations (ICLR ’14), 2014.

[5] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas and Martin Wattenberg, SmoothGrad: removing noise by adding noise, Workshop on Visualization for Deep Learning in ICML, 2017

[6] 中村鴻介, 山口実靖, “機械学習による主観文書分類結果の解釈性の付与に関する一考察”, WebDB Forum 2019, 1C-1, 2019.

[7] 中村鴻介, 山口実靖, “アテンションを用いた深層学習の分類結果に対する解釈性付与に関する一考察”, 情報処理学会 第83回全国大会, 6L-08, 2021.