文章間類似性判定モデルの根拠の抽出

中井香那子†　河田　友香††　山本　岳洋††

† 兵庫県立大学 社会情報科学部　〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
†† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †fa20h053@guh.u-hyogo.ac.jp, ††ad22l019@guh.u-hyogo.ac.jp,t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では，2 つの文の類似性を判定し,「類似」「非類似」という判定の根拠となる文節の抽出に取り組む．まず，類似度が与えられた 2 つの文章ペアのデータに，類似度に応じて「類似」または「非類似」というラベルを付与し,「類似」「非類似」を判定する BERT モデルを作成した．その後，各文を文節に分割し，文節の類似度が高い順に各文から 1 文節ずつを取り出した文節ペアを作成した．最後に，文節ペアをそれぞれマスクした状態で学習した BERT に入力し，類似に分類される確率をマスク前と比較することにより，変化の大きい文節ペアを判定に影響を与える根拠として抽出した．その結果，類似文からは類似度の高い文節を共通語として，非類似文からは類似度は高くないものの対応関係を持つ文節を差異として抽出することができた．また，誤分類に影響を与えた文節を抽出することにより，誤って学習している部分や学習が不十分であった部分を発見した．

キーワード　XAI，テキスト分類，BERT

1　は じ め に

インターネットの発展により，身の回りには情報があふれている．情報が簡単に手に入ることにより，文章を比較する機会やある文に関連する文章を探す機会は増加したと考えられる．例えば，オンラインショッピングでは 2 つの商品情報を比較し，商品の違う部分を見て購入する商品を決定する．具体的にイヤホンを購入する場面を考える．イヤホンの商品情報には音質やバッテリー，デザインといったカテゴリの情報が存在する．商品を検討する際は，まず同じカテゴリの情報を比較し，それぞれの商品の特徴を手に入れる．例えば音質について商品 A は「圧巻の**重低音**を再生する新規設計ドライバー．完全ワイヤレスで重要な装着感を追求しながら，圧巻の**重低音**を再生する大口径ドライバーを新規開発．高い密閉性により**低音**を逃しません.」，商品 B は「**高音域から低音域までバランスの良い**クリアな音質．**高音から低音までバランスが良く**，特にボーカルや楽器の音を自然でクリアにお楽しみいただけます.」という記載があったとする．このような説明文から商品 A は重低音重視であるのに対し，商品 B は高音から低音までのバランス重視であるという違いを手に入れる．また，低音重視といった手に入れた商品情報に関連するレビューを探すことで，より詳細な情報や他の人の感想を手に入れることもある．そして，手に入れた情報を総合して購入する商品を最終的に決定する．

以上のように，類似，関連している文章を比較する場面は日常生活で多くあり，類似性判定を行うモデルが作成されている．しかし，ディープラーニングを用いて類似性判定のモデルを作成すると人間には内部でどのような処理が行われているか分かりづらいという問題がある.「モデルからなぜこのような出力結果が得られるのかわからない」というブラックボックスモデルは問題視されており，BERT [6] では Attention [5] の可視化や LIME [13] が根拠の可視化として用いられている．こういったモデルの判断の根拠は，モデルの出力結果を理解しやすくなるだけでなく，モデルの性能の評価，モデルの精度の向上に使うことができる．例えば，正解ラベルと予測ラベルが一致しない場合，その根拠を抽出することでうまく学習が行えていない部分を発見し，性能の改善につなげることができる．また，正解ラベルと予測ラベルが一致している場合でも，正しい判断基準から判定を行えているか確認ができる．

さらに，根拠を可視化することができれば情報の比較の手助けになる．2 文の類似性判定においてモデルが人間と同様に，似ている単語の組み合わせに注目して類似と判定，似ていない単語の組み合わせに注目して非類似と判定しているならば，類似の根拠は 2 つの文の共通点，非類似の根拠は 2 つの文の相違点となる．そのような根拠が可視化できれば，商品を比較する際，長い文章をすべて読まずとも，2 つの商品の文章から一目で各カテゴリにおける共通点と相違点を確認できる．

既存研究では，2 つのラベルの分類においてそれぞれの分類の根拠の抽出を行うことで，各ラベルの特徴語を抽出する研究 [20] や，文書集合間の共通語，特徴語集合を可視化する研究 [21] が行われている．しかし，単語を抽出する際,「綺麗」と「汚い」のような各文の単語の対応関係には注目していない．

本研究では，類似度が与えられている 2 つの文章ペアのデータセットを用いて，類似性判定を行う BERT モデルを作成し，そのモデルの判定の根拠となる単語の抽出を行う．単語を抽出する際に，各文の文節の類似度を比較し，文節ペアを用いることにより，各文の単語の対応関係に注目した根拠の抽出を行った．同時に，モデルの性能の評価として誤分類の根拠を抽出し，モデルが上手く学習できていない部分を発見した．こういった根拠の抽出を商品の説明文に応用し，根拠を可視化することができれば，図 1 のように文章から共通点，相違点を探しやすくできるよう支援が行える．

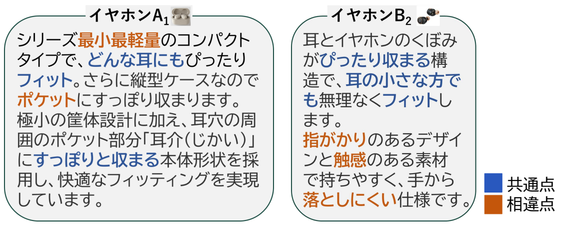

図 1　商品説明文における根拠の可視化の例.

2 関 連 研 究

2.1 類似性判定

単語や文章の類似性判定を利用するタスクとして，クエリと文書の適合性判定や同義語判定，類似文判定などが挙げられる．

クエリと文書の適合性判定のタスクの研究として，杉木らは商品を検索するクエリと商品レビューからそれぞれ要求と意見を抽出し，要求に合う商品の提示を行うシステムを開発している [17]．また，重要語に注目したレビューの要約の研究として，神谷らは，PageRank アルゴリズムを利用して抽出した重要語からクエリを生成し，クエリを用いて要約文を生成する方法を提案している [16]．

同義語判定のタスクとして，ある単語に対して前後の文をもとに適切な意味を判定するというような語義曖昧性解消問題がある．語義曖昧性解消の研究として，谷田部らは事前学習済み BERT を用いて用例文ペアにおける単語の同義判定を行っている [19]．さらに，同義関係を表したグラフと，グラフニューラルネットワークにより語義を判別するモデルの構築を行っている [18]．Mikolov らは単語のベクトル化の手法として Word2Vec という手法を提案している [10]．Pilehvar らは単語の意味の類似性を測定するために，単語の意味に対する確率的表現を使用した方法を提案している [11]．

類似文判定タスクの研究として，Batanović らはフレーズと文や，文と段落といった異なるレベルのテキスト間の類似度をはかる研究を行っている [1]．データセットにはセルビア語の新聞記事を用いて作成したデータセットを使用し，ボスニア語，クロアチア語，モンテネグロ語，セルビア語を用いて事前学習を行った BERTić モデルや，104 個の異なる言語を用いて事前学習を行った多言語 BERT モデルを利用して類似度をはかっている．また，Reimers らは 2 つの入力文の埋め込み表現を比較する BERT をファインチューニングしたモデルとして Sentence-BERT [12] を提案している．本研究では，2 つの文章の類似，非類似の判定を行う方法として BERT を，文節間のコサイン類似度をはかる方法として Sentence-BERT を用いる．

2.2 根拠や特徴語の抽出

XAI が注目されている現在，画像やテキストデータなど様々な分野で根拠推定の研究が行われている．

Lundberg らはゲーム理論のシャープレイ値を用いてモデルの解釈を行う SHAP という手法を提案している [9]．Ribeiro らは入力データの一部を隠すことによる出力の変化からモデルの解釈を行う LIME という手法を提案している [13]．Chen らはニューラルネットワークの解釈可能性を向上させるため，テキスト分類の階層的説明を生成する手法 [3] や，解釈可能性に加え，モデルの予測精度を向上させるため，単語マスク層を介してタスク固有の重要な単語を学習する VMASK という手法を提案している [2]．Kokalj らは BERT を含む Transformer モデルに対して SHAP を適応させる TransSHAP という手法を提案している [7]．Yan らは分類ラベルに関連付けられたトピックを分類の根拠として自動生成する分類器 HIMT を提案している [14]．Lei らはテキスト分類の根拠抽出として，ジェネレータとエンコーダを用いて入力テキストの一部を根拠として抽出するモデルの学習を行っている [8]．

テキストデータに対し LIME を用いて特徴語を抽出した研究として，小野川らのホテルのレビューを分類するモデルについて特徴語を抽出した研究が挙げられる [20]．コンフォートホテルとアパホテルのレビューを用いて，LSTM によりどちらのホテルのレビューかを分類した．そして分類結果の理由となる単語を LIME により単語を抽出したところ，各ホテルのレビューにおける単語出現頻度の差を表すカイ 2 乗値の上位の単語と一致した．さらに，カイ 2 乗値では抽出できなかったホテルの違いを表す特徴語を SP-LIME により抽出することができている．

また，Choudhary らは自然言語処理のブラックボックスモデルの解釈性についての調査を行っている [4]．局所的解釈可能性アプローチと，大局的解釈可能性アプローチに分けて分析している．局所的解釈可能性アプローチとは特定の入力に対する予測の根拠を示すことにより説明を行う方法である．一方で，大域的解釈可能性アプローチは複雑なモデルの全体を処理の簡単なわかりやすいモデルで表現することにより説明を行う方法である．それぞれの長所と短所を分析し，状況に応じて使い分ける必要があると結論付けている．

また，文書間の差異を表す特徴語の抽出として鷹田らは，文書集合間の共通語集合と特有語集合を可視化することによる理解支援の研究を行っている [21]．この研究では特有語を，他の文書集合から差別し，同時に関連性を示すような単語と定義し，対数尤度比とコサイン類似度を用いた文脈類似度によって抽出している．

本研究では，類似度の高い文節をペアとしてマスクすることによる類似に分類される確率の変化を見ることで，対応関係を持つ文節を類似性判定の根拠として抽出した．

3 文節ペアのマスクによる類似性判定モデルの根拠の抽出

本節では類似非類似の根拠となる文節の抽出手法について述べる．例文 A「スケートボードでジャンプしている男性がいます.」と例文 B「男性がスケートボードでジャンプしています.」

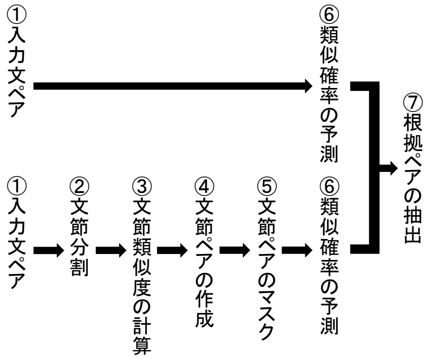

図 2　提案手法の概要.

という 2 文を用いて説明を行う.

3.1　提案手法の概要

提案手法の概要を図 2 に示し，以下で概要を説明する．図 2 の類似確率の予測を行うモデルに対し，入力文ペアをそのまま入力して得られた確率と，一部をマスクして入力して得られた確率を比較することで根拠を抽出する．以下に手順を示す.

（1）　入力文ペア

- 類似または非類似ラベルをつけた 2 文のペアを入力する.

（2）　分割文節

- 入力文をそれぞれ文節単位で分割する.

（3）　文節類似度の計算

- 各文から 1 つずつ文節を取り出し Sentence-BERT を用いて類似度を算出する.

（4）　文節ペアの作成

- 文節間の類似度の高い順に文節ペアを作成する.

（5）　文節ペアのマスク

- 文節ペアをマスクした状態で BERT に入力する.

（6）　類似確率の予測

- 予測分類確率を出力する BERT モデルの学習を行う.
- 予測分類確率とは類似に分類される確率と非類似に分類される確率である.
- 入力文ペアを類似性判定を行う BERT に入力する.

（7）　根拠ペアの抽出

- マスク前後で予測分類確率を比較する.
- 予測分類確率が逆転した文節ペアを根拠として抽出する.

また，文節ペアのマスク前後の予測分類確率の変化による根拠の抽出について，比較方法を図 3 に示す．図 3 では，例文の「スケートボードで」と「スケートボードで」という文節ペアをマスクする前と後の予測分類確率を比較している．マスクする前は類似に分類される確率が 0.9 であるのに対し，マスクした後は類似に分類される確率が 0.2 まで下がったとする．この場合，予測分類確率が逆転しているため,「スケートボードで」と「スケートボードで」という文節ペアは類似という判定に影響を与える根拠として抽出することができる．反対に,「男性が」と「男性が」という文節ペアをマスクすると，類似に分類される確率が 0.8 であったとする．この場合，マスク前も類似に分類される確率が 0.8 であり予測分類確率は変化していないため,「男性が」と「男性が」という文節ペアは類似という判定に影響を与えないと判断し，根拠として抽出しない.

3.2　類似性判定モデルの学習

本研究では BERT を用いて入力文ペアが類似または非類似である確率を予測するモデルを作成する．BERT とは双方向 Transformer Encoder によって構成される自然言語処理モデルである．ファインチューニングにより，分類や質問応答といった様々なタスクにおいて高精度な処理を行うことができる.

BERT には類似または非類似というラベルがつけられたペアとなっている文章を [SEP] トークンでつなぎ，2 文を同時に入力する．また，2 文を同時に入力するため，1 文目と 2 文目のトークンを判別するために Segment ID を用いる．Segment ID は 1 文目に 0，2 文目に 1 を割り当てる．出力は予測分類確率である．予測分類確率は [CLS] トークンのベクトルをシグモイド関数により変換することで算出した．入力と出力の例を図 4 に示す.

3.3　文節ペアの作成

本研究では，根拠の抽出の際文節ペアを利用する．２文の類似性判定を行うと，類似文には共通するような単語，非類似文には差異をあらわす単語というような対応関係を持つ単語が 2 文に存在すると考えられる．そこでペアを作成することにより対応する単語を根拠として同時に抽出することができる．また，単語単位での分割は助詞や助動詞などそれ単体では意味が理解しにくい単語が抽出される可能性があるため，意味を持つまとまりとして文節を 1 つの単位とする．文節の類似度を用いて文節同士のペアを作成する．このペアは入力文ペアの 1 文目から 1 文節，2 文目から 1 文節を選択したものであり，文節間の類似度の高いものからペアを作成する.

3.3.1　入力文ペアの文節分割

BERT に入力する 2 文をそれぞれ GiNZA により文節に分割する．例では文 A が「スケートボードで/ジャンプしている/男性が/います.」，文 B が「男性が/スケートボードで/ジャンプしています.」というように分割される.

3.3.2　文節の類似度比較

各文節を Sentence-BERT に入力し，文節間の類似度を求める．Sentence-BERT とは，2 つの文章をそれぞれ同一の BERT に入力して各文章のベクトルを生成し，生成したベクトルを用いて文章の関係性を判定する手法である．各文章についてベクトルを生成するため，大量の文章がある場合，高精度かつ高速に類似度を求めることができるという特徴がある．本研究では，文章ではなく文節のベクトルを生成し，類似度を算出する．具体的には，図 5 のように各文から文節を 1 つずつ選択して Sentence-BERT に入力し，コサイン類似度を算出する．まず，図 5 のように各文の 1 文節目を入力する．次に，文 A の 1 文節目の「スケートボードで」と文 B の 2 文節目の「スケートボードで」を入力するというように全ての文節の組み合わせでコサイン類似度を求める.

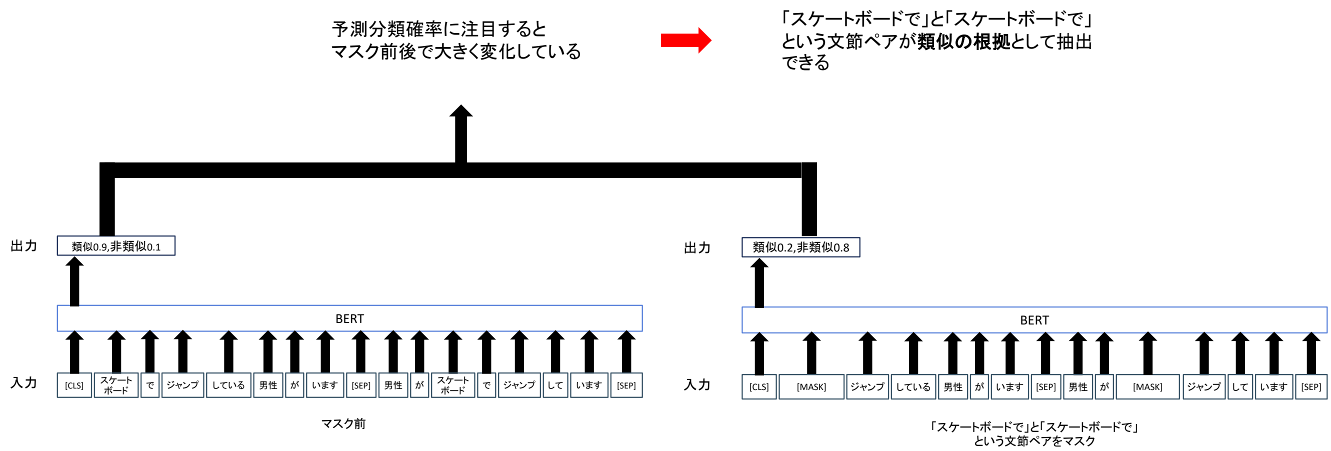

図 3　マスク前後の予測分類確率の変化による根拠の抽出.

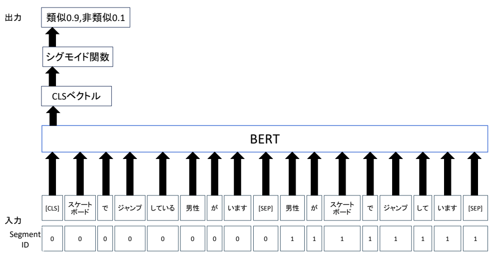

図 4　分類確率を予測する BERT モデルの入力と出力.

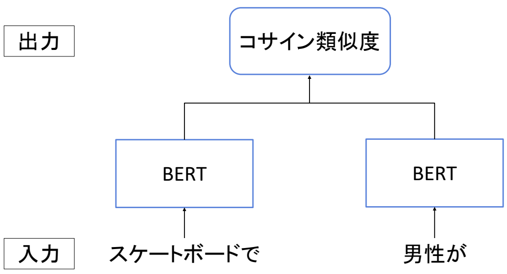

図 5　Sentence-BERT による文節の類似度比較.

3.3.3　文節ペアの作成

コサイン類似度の高いものから順番に，文節が重複しないよう対応付けを行い，文節のペアを作成する．仮に，例文における文節の類似度が高い上位 3 件が類似度の高い順に「スケートボードで」と「スケートボードで」,「スケートボードで」と「ジャンプしています.」,「ジャンプしている」と「ジャンプしています.」だったとする．この場合 1 つ目の文節ペアは 1 番類似度の高い「スケートボードで」と「スケートボードで」になる．次に類似度の高いペアは「スケートボードで」と「ジャンプしています.」だが,「スケートボードで」が 1 つ目の文節ペアに含まれるため文節ペアからは省略し，2 つ目の文節ペアとしては「ジャンプしている」と「ジャンプしています.」を選択する．

例文での文節ペアは実際には類似度の高い順に「スケートボードで」と「スケートボードで」,「男性が」と「男性が」,「ジャンプしている」と「ジャンプしています.」となる．

また，文節ペアは文節数の少ない文章の文節数に合わせ，ペアを作ることのできない文節については対応がないものとする．例では文 A が 4 文節，文 B が 3 文節に分割できる．文節数の少ない文章の文節数に合わせるため文節ペアは 3 つとなり，1 文目の「います.」という文節を対応なしとして扱う．

3.4　根拠の抽出

3.2 節で学習した類似性判定モデルに文節ペアをマスクした状態で 2 文を入力し予測分類確率の変化を観察する．例えば，例文のうち「ジャンプしている」と「ジャンプしています.」という文節ペアをマスクする場合，入力は「[CLS] スケートボードで [MASK] 男性がいます．[SEP] 男性がスケートボードで [MASK][SEP]」となる．そして，何もマスクしていない状態の予測分類確率と比較し，予測分類確率が逆転した場合，文節ペアを判定に影響を与える根拠として抽出する．また，対応なしとして扱う文節についても，その文節のみをマスクする．そして，マスク前後で予測分類確率が逆転した場合は根拠として抽出する．例では文 A の「います.」が対応なしとなったため,「います.」のみをマスクし,「[CLS] スケートボードでジャンプしている男性が [MASK][SEP] 男性がスケートボードでジャンプしています．[SEP]」という形で入力する．

4　実　　験

4.1　デ　ー　タ

研究に用いたデータについての説明を行う．早稲田大学とヤフー株式会社が構築・公開している JGLUE [15] の JSTS データセットを使用する．JGLUE とは日本語の言語理解モデル用

表 1　JSTS データセット.

類似度	文 A	文 B
0.0	ジャンプ台からスキーヤーがジャンプをしています.	茶色い猫の顔が，あおむけになっています.
1.0	男性が腕を組んでパソコンを見ています.	1 人の女性がベンチに座って本を読んでいます.
2.0	海上で女の子がサーフィンをしています.	男の子が海で遊んでいます.
3.0	柵で囲まれた芝生と道路が見えます.	柵で囲まれた池と道路が見えます.
4.0	男性が子供を抱き上げて立っています.	坊主頭の男性が子供を抱いて立っています.
5.0	真っ赤な二階建てのバスが停まっています.	赤い二階建てのバスが止まっています.

のベンチマークであり，中でも JSTS は文ペアの類似度推定タスク用に構築されている．具体的には，JSTS は意味が完全に異なるものを 0，意味が等価であるものを 5 として 0〜5 までの数字が 0.2 単位で類似度として付与されている．データセットの一部を表 1 に示す．類似度は文 A と文 B の類似度を表す．本研究では，類似度が 3.8 以上 5.0 以下であるデータに類似，類似度が 2.0 以上 2.6 以下までのデータに非類似とラベルを付ける．実際に文の比較を行う際，例えばイヤホンの購入を考えている場面では，デザインに関する 2 つの文や音質に関する 2 つの文など関連性を持つ 2 つの文を比較する．しかし，JSTS データの類似度 0.0 のデータは「ジャンプ台からスキーヤーがジャンプをしています.」と「茶色い猫の顔が，仰向けになっています.」のように関連性を持たない組み合わせとなっている．一方で類似度 2.0 のデータは「海上で女の子がサーフィンをしています.」と「男の子が海で遊んでいます.」のように意味は全く違うが，海という点で関連性を持っている．そこで，非類似のなかでも 2.0 未満のペアについては 2 つの文章間に関連性がないものとし，意味は似ていないが関連性を持つ 2.0 以上 2.6 以下のデータに「非類似」ラベルを付与する．また，類似度 4.0 や 5.0 のデータは「男性が子供を抱き上げて立っています.」と「坊主頭の男性が子供を抱いています.」のように文は一致しないものの一部言葉を付け加えたり，言い換えたりした意味が類似しているペアとなっている．そこで，4.0 以上のデータに「類似」ラベルを付与しようとしたところ，データ数が少なかったため，類似度の幅を広く 3.8 以上 5.0 以下と設定した．データ数は訓練データ 3,200 件，検証データ 400 件，テストデータ 400 件の計 4,000 件である．

また，根拠を抽出するにあたって，分析の対象はテストデータのうち未知語を含まない 2 文の組み合わせのみを利用した．

4.2 実験設定

類似性判定モデルの作成には東北大学の学習済み BERT モデル [1]を用いた．また，学習率は 2.0×10^{-5}，patience は 3，バッチサイズは 16，エポック数は 10，入力の最大長は 512 と設定した．また，文節の類似度の測定には日本語用の事前学習済み Sentence-BERT モデル [2]を用いた．

4.3 比較手法

類似，非類似の根拠の抽出の比較手法として LIME を用いる．LIME とはまず学習したモデルに文章の一部をマスクしたものを入力する．次に，一部をマスクした入力とその入力に対して得られた類似または非類似のラベルをペアとして線形回帰を行う．そして，線形回帰の結果から出力に大きな影響を与えた単語を特徴語として抽出する．比較を行う際には，回帰に使用する一部をマスクした入力とそれに対する出力のペアデータの数を出力結果が安定する 5,000 件とし，上位 10 件の単語を抽出した．また可視化の際には，類似の根拠を青色，非類似の根拠をオレンジ色で示す．

5 結　　果

5.1 類似文ペアを入力した際の根拠の抽出例

正解ラベルが「類似」であり，予測分類確率も「類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「非類似」によるような文節ペアを「類似」の根拠として抽出した．

具体例として,「スケートボードでジャンプしている男性がいます.」と「男性がスケートボードでジャンプしています.」という類似文について抽出した結果を示す．この 2 文はそれぞれ「スケートボードで/ジャンプしている/男性が/います.」「男性が/スケートボードで/ジャンプしています.」という文節に分割される．さらに，各文節の類似度を測り，類似度の高いものから文節ペアを作ると「スケートボードで」と「スケートボードで」,「男性が」と「男性が」,「ジャンプしている」と「ジャンプしています.」となった．また，1 文目が 4 文節，2 文目が 3 文節に分割される．文節数の少ない文章の文節数に合わせるため，文節ペア数は 3 つとなり，1 文目の「います.」という文節は対応なしとして扱う．マスクをしていない元の文と，各文節ペアをマスクした文の予測分類確率は表 2 のようになった．表は上からマスク前，文節類似度の高い順にペアのマスク後，対応なしとなった文節のマスク後の順である．

表 2 より「スケートボードで」や「男性が」といった文節は 2 つの文に共通する文節だが，判定には大きな影響を与えていないことが分かる．一方で,「ジャンプしている」と「ジャンプしています.」という文節ペアは，予測分類確率が逆転しており，類似という判定に大きな影響を与えていることが分かる．また，対応のなかった「います.」は 1 文目のこの文節のみをマスクしたところ，予測分類確率にはほとんど変化がなかったことから，判定に大きな影響を与える文節ではないといえる．

これらの結果より，この 2 文において類似の判定の根拠は「ジャンプしている」と「ジャンプしています.」という文節ペアだといえる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結

1：https://github.com/cl-tohoku/bert-base-japanese-v3

2：https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

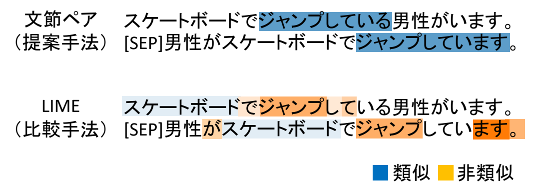

図 6　類似文から抽出された根拠の比較.

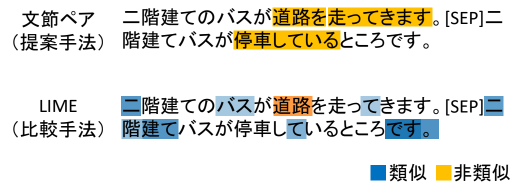

図 7　非類似文から抽出された根拠の比較.

果と提案手法による根拠を抽出した結果の比較を図 6 に示す.「スケートボード」という 2 つの文両方に出現する単語が類似の根拠として抽出できた一方で,「ジャンプ」という共通語は非類似の根拠として抽出された．また，非類似の根拠となる単語が多く抽出されている点や，句読点や助詞，助動詞が抽出されている点から，類似の根拠となる共通部分はわかりづらい結果となっている.

5.2　非類似文ペアを入力した際の根拠の抽出例

正解ラベルが「非類似」であり，予測分類確率も「非類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「類似」によるような文節ペアを「非類似」の根拠として抽出した.

具体例として,「二階建てのバスが道路を走ってきます.」と「二階建てバスが停車しているところです.」という非類似文について抽出した結果を示す．2 文はそれぞれ「二階建ての/バスが/道路を/走って/きます.」「二階建てバスが/停車している/ところです.」という文節に分割される．さらに，類似度の高いものから文節ペアを作ると「バスが」と「二階建てバスが」,「きます.」と「ところです.」,「走って」と「停車している」となる．また，1 文目は 5 文節，2 文目は 3 文節に分割できる．文節数の少ない文章の文節数に合わせるため文節ペアは 3 つとなり，1 文目の「二階建ての」「道路を」という文節を対応なしとして扱う．マスク前と各文節をマスクした予測分類確率の結果を表 3 に示す.

表 3 より「バスが」と「二階建てバスが」という文節ペアは予測分類確率に大きな変化はないことがわかる．一方で,「きます」と「ところです」という文節ペアは予測分類確率が逆転しないものの，15 %程度類似に分類される確率が上がっており，非類似という判定に影響を与えていると考えられる．さらに,「走って」と「停車している」という文節ペアは予測分類確率が逆転しており，判定に大きな影響を与える非類似の根拠といえる．この 2 文節の特徴として，文節間の類似度は 0.3411 となっており他の文節ペアに比べると類似度が低いことが挙げられる．ただし,「走る」「停車する」というバスの動きを表す語として関連性は持っているといえる.

対応関係のない文節については 1 文目の 1 文節のみをマスクして比較を行った．その結果,「二階建ての」は予測分類確率に変化が見られなかったが,「道路を」は予測分類確率が逆転した.「二階建ての」は文節としては対応がないが，2 文目の「二階建てバス」という文節に「二階建て」という語が含まれるため判定に影響を与えないと考えられる.

これらの結果より，類似度は低いものの対応関係を持つ「走って」「停車している」という文節ペアと片方のみに出現する「道路を」という文節が非類似文の差異として適切に抽出できたといえる.

比較対象として同じ文に LIME を適用し，根拠を抽出した結果と提案手法による根拠を抽出した結果の比較を図 7 に示す. LIME では「道路」という 1 文目にのみ出現する語が非類似の根拠として抽出された一方で,「走る」と「停車」という対になっている語は抽出できなかった．また,「二階建て」という語の一部である「二」のみが抽出されたり,「して」や「です」などその単語単体で意味の通じない単語が抽出されたりし，非類似の根拠としてわかりづらい結果となった.

5.3　モデルが誤分類を行った際の根拠の抽出例

正解ラベルは「類似」だが予測分類確率は「非類似」である確率が高くなったり，正解ラベルは「非類似」だが予測分類確率は「類似」である確率が高くなったりしたような誤分類においても，正しく類似，非類似と判定された文章と同様に文節ペアを作成し，予測分類確率をマスク前後で比較した．そして，予測分類確率がマスク前後で逆転している文節ペアを誤った分類へ大きな影響を与えたものとして，学習が不足していた部分や間違って学習していた部分の発見につなげた.

まず，正解ラベルが「類似」であったが，予測分類確率は「非類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「類似」によるような文節ペアを，誤って非類似と判定した根拠として抽出した．その結果，2 つの文の差異となる部分が非類似という判定に影響を与えていることが分かった.

具体例として「白のジャンボジェット機が青空を飛んでいます.」「青い空を，飛行機が飛んでいます.」という 2 つの文の抽出結果を示す．2 文はそれぞれ「白の/ジャンボジェット機が/青空を/飛んでいます.」「青い/空を，/飛行機が/飛んでいます.」という文節に分割された．さらに，類似度の高い順に「飛んでいます.」と「飛んでいます.」,「ジャンボジェット機が」と「飛行機が」,「青空を」と「空を,」,「白の」と「青い」という文節ペアになった．マスク前と各文節をマスクした予測分類確率の結果を表 4 に示す．表 4 より「飛んでいます.」と「飛んでいます.」や「青空を」と「空を,」といった文節ペアは予測分類確率に大きな変化はないことがわかる．一方で,「白の」と「青い」という文節ペアは予測分類確率が逆転しており，非類似という

表 2　類似文のマスクと予測分類確率.

文 A	文 B	類似	非類似
スケートボードでジャンプしている男性がいます.	男性がスケートボードでジャンプしています.	0.9997	0.0002
[MASK] ジャンプしている男性がいます.	男性が [MASK] ジャンプしています.	0.9960	0.0039
スケートボードでジャンプしている [MASK] います.	[MASK] スケートボードでジャンプしています.	0.9238	0.0769
スケートボードで [MASK] 男性がいます.	男性がスケートボードで [MASK]	0.0004	0.9995
スケートボードでジャンプしている男性が [MASK]	男性がスケートボードでジャンプしています.	0.9991	0.0008

表 3　非類似文のマスクと予測分類確率.

文 A	文 B	類似	非類似
二階建てのバスが道路を走ってきます.	二階建てバスが停車しているところです.	0.0898	0.9101
二階建ての [MASK] 道路を走ってきます.	[MASK] 停車しているところです.	0.0011	0.9988
二階建てのバスが道路を走って [MASK]	二階建てバスが停車している [MASK]	0.0011	0.7740
二階建てのバスが道路を [MASK] きます.	二階建てバスが [MASK] ところです.	0.6139	0.3860
[MASK] バスが道路を走ってきます.	二階建てバスが停車しているところです.	0.0007	0.9992
二階建てのバスが [MASK] 走ってきます.	二階建てバスが停車しているところです.	0.9655	0.0344

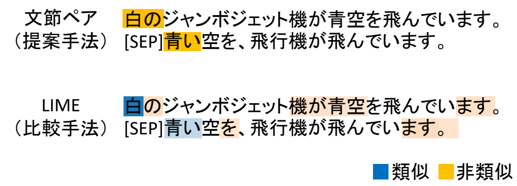

図 8　非類似を類似と判定した入力文ペアから抽出された根拠の比較.

誤分類に影響を与えた文節ペアだといえる．さらに,「白の」と「青い」は文節類似度が低く，それぞれを単体でマスクしたところ,「青い」は非類似に分類される確率が 0.9932 と他の文節ペアをマスクしたときと変化が見られなかったが,「白の」は非類似に分類される確率が 0.1590 と予測分類確率が逆転した．このことから，この 2 文の判定を行う際に,「白の」という 2 文の差異に注目しすぎた結果，非類似という誤分類につながったと考えられる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結果と提案手法による根拠を抽出した結果の比較を図 8 に示す．「青い」という単語が類似の根拠として抽出された一方で「青空」は非類似の根拠として抽出されるなど対応関係のあるものが別々の根拠として示された．また,「白」という一文目のみに出現する単語が類似の根拠として抽出されたほか，上位 10 件のなかに付属語が多く見られた．

次に，正解ラベルが「非類似」であったが，予測分類確率は「類似」である確率が高くなった 2 文について，文節ペアをマスクすることにより，判定が「非類似」によるような文節ペアを，誤って類似と判定した根拠として抽出した．その結果，2 つの文に共通する部分が類似という判定に影響を与えていることが分かった．

具体例として「芝生の真ん中に，消火栓が設置されています.」「道路の真ん中に，消火栓が設置されています.」という 2 つの文の結果を示す．2 文はそれぞれ「芝生の/真ん中に，/消火栓が/設置されています.」「道路の/真ん中に，/消火栓が/設置されています.」という文節に分割された．さらに，類似度の高い順に「消火栓が」と「消火栓が」,「設置されています.」と「設置されています.」,「真ん中に,」と「真ん中に,」,「芝生の」と「道路の」という文節ペアになった．マスク前と各文節をマスクした後の予測分類確率の結果を表 5 に示す．表 5 より「消火栓が」と「消火栓が」や「芝生の」と「道路の」といった文節ペアは予測分類確率に大きな変化はないことがわかる．一方で,「設置されています.」と「設置されています.」や「真ん中に,」と「真ん中に,」という文節ペアは予測分類確率が逆転しており，類似という誤分類に影響を与えた文節ペアだといえる．このことから，この 2 文の判定を行う際に,「設置されています.」や「真ん中に,」といった 2 文に共通する文節に注目した結果，類似という誤分類につながったと考えられる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結果と提案手法による根拠を抽出した結果の比較を図 9 に示す．「真ん中」や「消火栓」という 2 つの文両方に出現する単語が非類似の根拠として抽出された．また，全体としては類似と判定しているにもかかわらず，上位 10 件は非類似の根拠に偏った結果となった．

5.4　根拠が抽出できなかった例

文節ペアをマスクすることにより，根拠が上手く抽出できないものもあった．具体例として「綺麗に掃除されたトイレがあります.」「茶系統のタイル張りの部屋のトイレです.」という 2 つの文の抽出結果を示す．2 文はそれぞれ「綺麗に/掃除され

表 4　類似を非類似と判定した文のマスクと予測分類確率.

文 A	文 B	類似	非類似
白のジャンボジェット機が青空を飛んでいます.	青い空を，飛行機が飛んでいます.	0.0007	0.9992
白のジャンボジェット機が青空を [MASK]	青い空を，飛行機が [MASK]	0.0005	0.9995
白の [MASK] 青空を飛んでいます.	青い空を，[MASK] 飛んでいます.	0.0018	0.9981
白のジャンボジェット機が [MASK] を飛んでいます.	青い [MASK] 飛行機が飛んでいます.	0.0004	0.9996
[MASK] ジャンボジェット機が青空を飛んでいます.	[MASK] 空を，飛行機が飛んでいます.	0.9995	0.0005

表 5　非類似を類似と判定した文のマスクと予測分類確率.

文 A	文 B	類似	非類似
芝生の真ん中に，消火栓が設置されています.	道路の真ん中に，消火栓が設置されています.	0.9992	0.0007
芝生の真ん中に，[MASK] 設置されています.	道路の真ん中に，[MASK] 設置されています.	0.9615	0.0384
芝生の真ん中に，消火栓が [MASK]	道路の真ん中に，消火栓が [MASK]	0.2881	0.7118
芝生の [MASK] 消火栓が設置されています.	道路の [MASK] 消火栓が設置されています.	0.3990	0.6009
[MASK] 真ん中に，消火栓が設置されています.	[MASK] 真ん中に，消火栓が設置されています.	0.9993	0.0006

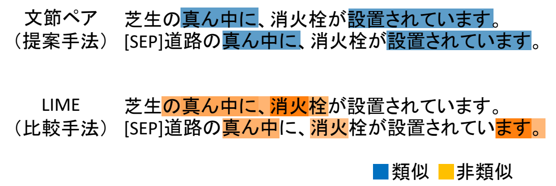

図 9　非類似を類似と判定した入力文ペアから抽出された根拠の比較.

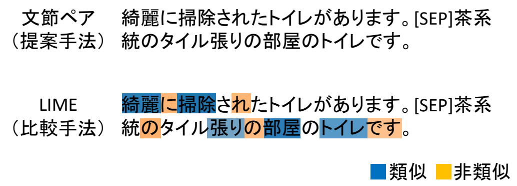

図 10　提案手法では根拠が抽出できなかった例.

た/トイレが/あります.」「茶系統の/タイル張りの/部屋の/トイレです.」という文節に分割された．さらに，類似度の高い順に「トイレが」と「トイレです.」,「あります.」と「部屋の」,「掃除された」と「タイル張りの」,「綺麗に」と「茶系統の」という文節ペアになった．マスク前と各文節をマスクした後の予測分類確率の結果を表 6 に示す．表 6 よりどの文節ペアでも予測分類確率に大きな変化はないことがわかる．これらの文節ペアの特徴として「トイレが」と「トイレです」を除く 3 ペアは類似度 0.4 を下回っており，2 つの文節間に関連がないことが挙げられる．このような結果から，文節同士が関連を持っていなければ文節ペアを根拠として抽出することは難しいといえる．

比較対象として同じ文に LIME を適用し，根拠を抽出した結果を図 10 に示す.「トイレ」という 2 文に共通する単語が類似の根拠で抽出された一方で,「掃除」や「タイル」など片方の文にしか出現しない単語が類似の根拠として抽出された．また，非類似の根拠はすべて「です」や「に」など助動詞や助詞となり 2 文の違いは分かりづらい結果となった．

6　まとめと今後の課題

本研究では，各文の文節を用いてマスクを行い，予測分類確率を比較することで，類似文からは共通する部分，非類似文からは差異を表す部分を根拠として抽出することに取り組んだ．同時に，誤分類の根拠を抽出することで，学習の不足部分を検討した．

比較手法である LIME と比べて，LIME では正確に抽出することのできなかった共通点や，対応関係を持つ相違点を抽出することができた．また，文章の分割の単位として単語ではなく文節を採用することにより，助詞や助動詞単体ではなく意味を持つ言葉のまとまりを抽出することができた．

一方で，LIME では類似文の共通語に加えて差異を表す語，非類似文の差異を表す語に加えて共通語も抽出できたが，提案手法による抽出結果は類似文からは共通する文節のみ，非類似文からは差異を表す文節のみとなった．今後，類似文中の差異を表す語，非類似文中の共通語の抽出の方法を検討する必要がある．

本研究の限界点として，文章同士の対応関係がなければ根拠をうまく抽出できないという点が挙げられる．具体的には，非類似と判定される文章のうち「綺麗に掃除されたトイレがあります.」と「茶系統のタイル張りの部屋のトイレです.」では「トイレが」と「トイレです」という文節ペアを除き，文節間の対応関係がなく，文節ペアをマスクすることによる予測分類確率の変化は見られなかった．

表 6 根拠を抽出できなかった文のマスクと予測分類確率.

文 A	文 B	類似	非類似
綺麗に掃除されたトイレがあります.	茶系統のタイル張りの部屋のトイレです.	0.0007	0.9993
綺麗に掃除された [MASK] あります.	茶系統のタイル張りの部屋の [MASK]	0.0007	0.9993
綺麗に掃除されたトイレが [MASK]	茶系統のタイル張りの [MASK] トイレです.	0.0006	0.9994
綺麗に [MASK] トイレがあります.	茶系統の [MASK] 部屋のトイレです.	0.0005	0.9995
[MASK] 掃除されたトイレがあります.	[MASK] タイル張りの部屋のトイレです.	0.0007	0.9993

また，ペアを作成するにあたって文節を用いたため，ある文では 1 文節でまとまるものが，もう一方では 2 文節に分かれてしまいうまく比較できないということがあった．例えば，非類似で例を挙げた「二階建てのバス」と「二階建てバス」は同じ意味を表す語だが，前者は「二階建ての/バス」の 2 文節，後者は「二階建てバス」の 1 文節である．その結果文節ペアは「バス」と「二階建てバス」となり,「二階建てのバス」と「二階建てバス」という 2 つの語全体の比較を行うことはできなかった.

上記 2 つの問題点より，今後の課題として，文節同士の類似度だけでなく，文節に含まれる単語の品詞の組み合わせなどからペアを決めるといった他のペアの作成方法を検討することが考えられる.

また，本研究では短い文を用いて根拠の抽出を行ったが，実際オンラインショッピングでのレビューや商品説明文を比較することを考えると一文がより長くなることや，一文同士の比較だけでなく，複数の文章を比較することが考えられる．本研究から，文の一部分を使うことにより適切に根拠を抽出することができることが分かったが，今後，より長い文の比較に対応するためには，文節よりも適した単位でのペアの作成方法を検討する必要性がある.

謝辞 本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775, JP22H03905, による助成を受けたものです．ここに記して謝意を表します.

文 献

[1] Vuk Batanović and Maja Miličević Petrović. Cross-Level Semantic Similarity for Serbian Newswire Texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1691–1699, 2022.

[2] Hanjie Chen and Yangfeng Ji. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4236–4251, 2020.

[3] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5578–5593, 2020.

[4] Shivani Choudhary, Chatterjee Niladri, and Subir Kumar Saha. Interpretation of black box NLP models;a survey. *arXiv:2203.17081*, 2022.

[5] 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, 2019.

[6] 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*, pp. 4171–4186, 2019.

[7] Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pp. 16–21, 2021.

[8] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016.

[9] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 4768–4777, 2017.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*, 2013.

[11] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1341–1351, 2013.

[12] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You":Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, 2016.

[14] Hanqi Yan, Lin Gui, and Yulan He. Hierarchical Interpretation of Neural Text Classification. *Computational Linguistics*, Vol. 48, No. 4, pp. 987–1020, 2022.

[15] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. 言語処理学会第 28 回年次大会発表論文集, pp. 2023–2028, 2022.

[16] 神谷賢太郎, 原寛紀, 青山幹雄. 深層学習によるレビュー内の重要語に着目した要約方法の提案と評価. 第 82 回全国大会講演論文集, pp. 463–464, 2020.

[17] 杉木健二, 松原茂樹. 消費者の意見に基づく商品検索. 情報処理学会論文誌, Vol. 49, No. 7, pp. 2598–2603, 2008.

[18] 梨恵谷田部, 稔佐々木. 用例文間の意味的な類似関係を用いた半教師あり語義曖昧性解消. 情報処理学会論文誌, Vol. 62, No. 10, pp. 1724–1736, 2021.

[19] 谷田部梨恵, 佐々木稔. BERT の学習済みモデルを用いた用例文ペアの同義判定. 言語処理学会 第 26 回年次大会 発表論文集, pp. 824–827, 2019.
[20] 小野川稜之, 折原良平, 清雄一, 田原康之, 大須賀昭彦. 機械学習モデルの解釈手法による競合サービスと比較したレビュー分析. 日本ソフトウェア科学会大会論文集, Vol. 36, pp. 337–343, 2019.
[21] 鷹田和弘, 大澤幸生. 複数文書の相対的特徴可視化による理解支援. 人工知能学会第二種研究会資料, pp. 41–46, 2013.