DEIM Forum 2023 1a-1-3

Attentionに着目したストップワードの自動生成

桑原　悠希†　　鈴木　　優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1112 岐阜県岐阜市柳戸 1–1
E-mail: †y3033061@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし　システム開発者は，文書分類や検索の精度向上を目的として，ストップワードを使用することがある．ストップワードを適切に設定することによって，文書分類の精度が向上すると考えられる．しかし，BERT を用いた文書分類タスクにおいて，公開されている既存のストップワードリストは分類精度を向上させることに有効ではない．本研究では，分類を行った際の Attention に着目して，ストップワードの自動生成を行うシステムを構築した．また，自動で生成したストップワードの有効性を確かめるために実験を行った．実験の結果，自動で生成したストップワードを用いて分類を行った場合に，ストップワードを使用せずに分類を行った場合と比較して，精度の改善が見られる場合もあった．正解ラベルと異なるラベルが予測された文書の Attention のみに着目した場合よりも，正解ラベルと同じラベルが予測された文書との Attention の差に着目した場合の方が高い精度で分類できることが分かった．

キーワード　ストップワード，テキスト分類，Attention，BERT，機械学習，自然言語処理

1　はじめに

ストップワードとは，文書分類や検索を行う際に処理の対象外とする単語のことである [1]．システム開発者は，文書分類や検索の精度向上を目的として，ストップワードを使用することがある．ストップワードを適切に設定することによって，分類精度が向上すると考えられる．しかし，適切でないストップワードを使用した場合には，分類精度に悪影響を与えてしまうことがある [2]．

BERT [3] を用いた文書分類タスクにおいて，公開されている既存のストップワードリストは分類精度の向上に有効ではないことが分かっている [4]．そのため，分類精度の向上に有効なストップワードを作成する必要があると考えられる．

BERT には，Attention 機構 [5] が用いられている．Attention 機構とは，入力データのどの部分に注目すべきか学習する仕組みのことである．分類器が分類を行った際の Attention に着目することによって，分類器の判断根拠を解釈することができる．分類時の判断根拠となる部分から分類性能に悪影響を与えている部分を見つけられないかと考えた．そのため，BERT を用いて文書分類を行った際の Attention に着目した．

正解ラベルと同じラベルが予測されたデータを正解データ，正解ラベルと異なるラベルが予測されたデータを不正解データとする．不正解データ中で Attention の高い単語は，分類器が誤った予測をする要因になっていると考えられる．そのため，不正解データの Attention に着目してストップワードを生成することによって，分類精度を向上させることができると考えた．しかし，不正解データの Attention のみに着目すると，正解データに出現した場合にも Attention の高い単語がストップワードになることが考えられる．正解データにおいて Attention が高い単語は，分類器が正しい予測をすることに貢献していると考えられる．そのため，不正解データの Attention のみに着目してストップワードの生成を行うことは適切ではないと考えた．正解データで Attention の低い単語は，分類器が正しい予測をすることに貢献していないと考えられる．不正解データで Attention が高く，正解データで Attention の低い単語をストップワードとすることが良いと考えた．そのため，単語が不正解データに出現した場合と正解データに出現した場合との Attention の差に着目した．

ストップワードの生成方法を 2 種類提案する．2 種類の手法は同時には使用せず，ストップワードの除去を行う場合にはどちらか片方の手法のみで行う．一つ目は，単語が不正解データに出現した場合と正解データに出現した場合との Attention の差を，ストップワードリストに含まれる確率として扱う手法である．システム開発者は，ストップワードリストに含まれる単語を入力文書中から削除するという方法で，ストップワードの除去を行うことがある．しかし，既存のストップワードリストは分類精度の向上に有効ではない．そのため，分類精度の向上に有効なストップワードリストを新たに作成する必要があると考えた．作成したストップワードリストに含まれる単語を全て入力文書中から削除する．二つ目は，単語が不正解データに出現した場合と正解データに出現した場合との Attention の差を単語が削除される確率として扱う手法である．Attention の差が大きい単語は分類精度に悪影響を与えていると考えられる．しかし，Attention の差が大きい単語が必ずしも分類精度に悪影響を与えているとは限らない．そのため，Attention の差が大きい単語であっても，入力文書中から全て削除するのではなく，一部を削除しないで入力文書中に残す方法を考えた．入力文書中の単語一つ一つに対して，入力文書中から削除するかどうかを確率的に決める．一つ目の手法では，ストップワードリストに含まれる単語を全て削除するが，二つ目の手法では，同じ単語であっても削除する場合と削除しない場合がある．

提案手法を用いて生成したストップワードの有効性を確かめるために実験を行った．ストップワードを用いて BERT で文

書分類を行い，ストップワードを使用せずに分類を行った場合と精度の比較を行った．さらに，ストップワードの除去による精度の変化に有意な差があるかどうか確認するために統計的検定を行った．使用するデータセットの種類や分類の目的によって最適なストップワードは異なると考えられる．そこで，同一のデータセットを用いて目的の異なる分類を行った．また，3種類のデータセットを使用して実験を行った．

実験の結果，提案手法を用いて生成したストップワードを使用することによって精度の改善が見られる場合もあった．不正解データにおける Attention のみに着目した場合よりも，正解データにおける Attention との差に着目した場合の方が高い精度で分類できることが分かった．

本論文の貢献は以下の通りである．

- BERT を用いた文書分類タスクにおいて，精度の改善が見られることもあるストップワードの生成ができた．
- 不正解データにおける Attention のみに着目した場合よりも，正解データにおける Attention との差に着目した場合の方が高い精度で分類できることが分かった．

2 関連研究

ストップワードの生成についての研究が行われている．國府ら [6] は，テキストの内容推測を目的としたキーワード抽出タスクにおいて有効なストップワードの生成を行った．出現頻度の高い単語を何かしらの基準で選別するという方法でキーワード抽出をした．単語を選別するための基準にストップワードを用いた．ストップワードとして除去する対象は,「非語」「非内容語」「低内容語」の 3 種類を設定した.「非語」は句読点や記号を対象にした.「非内容語」は内容語ではない単語であり，機能語と呼ばれる単語を対象にした．機能語とは，単語間や文と文の文法的な関係性を示すのに用いられる単語である．品詞情報を用いて非内容語の除去を行った.「低内容語」は品詞情報を用いて単語の除去を行っても削除されない単語のうち，内容の推測に貢献しそうにない単語である．國府らは「低内容語」のリストの作成をした．作成したストップワードリストがキーワード抽出タスクにおいて有効に機能することが分かった．

Saiyed ら [7] は，ストップワードの生成を行った．ストップワードを作成する手法は 2 種類ある．一つ目は，手動でストップワードリストを作成しすべての単語をストップワードリストと照合する方法である．二つ目は，自動的にストップワードリストを作成する方法である．前者を静的手法，後者を動的手法とした．単語の出現頻度がジップの法則に従うことに着目した．出現頻度の上位と下位で閾値を決めてストップワードとなる単語を決めた．静的手法で 44.53 %、動的手法で 52.53 %文書のサイズを削減することができた．動的手法ではストップワードとすべきでないと考えられる単語もストップワードになってしまうことがあることが確認された．

ニューラルネットワークを用いた分類手法にストップワードの考え方を適用した研究が行われている．木村ら [8] は，ストップフレーズが文書分類タスクに与える影響の調査を行った．複数のサブワードで構成されるサブワード列をサブワードフレーズとした．サブワードの出現頻度はジップの法則に従う．出現頻度の高い単語は，文書の意味を表さない機能語と呼ばれる単語が多い．そのため，出現頻度の高い単語をストップワードとするという考え方がある [9]．このストップワードの考え方を用いて，出現頻度の高いサブワードフレーズのをストップフレーズとした．ストップフレーズをトークナイザの語彙に追加して行う実験と，ストップフレーズの抽出と文書分類を行うマルチタスク学習での実験を行った．実験の結果，ストップフレーズを考慮することによって分類精度が向上した．

國府らは，出現頻度の高い低内容語をストップワードとした．Saiyed らは，出現頻度の上位と下位から閾値を決めてストップワードとなる単語を決めた．木村らは，出現頻度の高いサブワードフレーズのをストップフレーズとした．本研究では，BERT を用いて文書分類を行った際の Attention に着目してストップワードを生成した．

3 提案手法

BERT を使用してテキストデータの分類を行った際に，正解ラベルと同じラベルが予測されたテキストデータを正解データ，正解ラベルと異なるラベルが予測されたテキストデータを不正解データとする．我々は，入力文書中に出現する単語が，正解データに出現した場合と不正解データに出現した場合との Attention の差に着目した．Attention の差を単語がストップワードになる確率として用いてストップワードの生成を行った．入力はテキストデータである．出力は単語と Attention の差のリストである．以下にストップワード生成の手順を示す．

（1） BERT を使用してストップワードの除去を行っていないテキストデータの分類を行う．

（2） 入力文書中の各単語が正解データに出現した場合の Attention の平均と不正解データに出現した場合の Attention の平均を求める．

（3） （2）で求めた単語の Attention の平均を用いて，単語ごとに不正解データに出現した場合と正解データに出現した場合との Attention の平均の差を求める．

（4） 出力として，単語と（3）で求めた Attention の差が保存されたリストが得られる．Attention の差を確率として扱い，以下に示す 2 種類の方法でストップワードの除去を行う．

- 手法 1：ストップワードリスト

（a） 確率をもとに，単語がストップワードになるかどうかを決める．

（b） ストップワードとなる単語をストップワードリストに追加する．

（c） 作成したストップワードリストに含まれる単語を，入力文書中から全て削除する．

- 手法 2：確率的ランダム除去

（a） 入力文書中に含まれる単語一つ一つについて，確率をもとに削除するかどうかを判定する．

（b） 削除すると判定された単語を入力文書中から削除する．

3.1 Attention

Attention 機構とは，入力データのどの部分に注目すべきか学習する仕組みのことである．Attention 機構は BERT にも用いられている．分類器が分類を行った際の Attention に着目することによって，分類器の判断根拠を解釈することができる．分類時の判断根拠となる部分から分類性能に悪影響を与えている部分を見つけられないかと考えた．そのため，BERT を用いて文書分類を行った際の Attention に着目した．

入力文書中に出現するそれぞれの単語について，不正解データに出現した場合の Attention の値と正解データに出現した場合の Attention の値にそれぞれ着目し, 単語ごとに Attention の値の平均を算出した．不正解データの文書集合を D，D に含まれる文書の総数を N とする．D に含まれる文書のうち，i 番目の文書を d_i とする．このとき i の値の範囲は，1 から N までである．d_i における，ある単語 w の出現する回数を n_i とする．D における，単語 w の出現する回数を n_i とする．d_i に出現する w のうち，j 番目に出現する w を w_j とする．このとき j の値の範囲は，1 から n_i までである．d_i に出現する w_j の Attention の値を $a(d_i, w_j)$ とする．正解ラベルと異なるラベルが予測された文書に出現した場合の単語 w の Attention の値の平均 $a_m(w)$ を以下の式で算出した．

$$a_m(w) = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{j=1}^{n_i} a(d_i, w_j)$$

同様の方法で，正解データに出現した場合の Attention の値の平均 $a_c(w)$ も求めた．

不正解データでの Attention$a_m(w)$ の値が高くなっている単語は，分類器が誤った予測をした際に注目していた単語であると解釈できる．不正解データでの Attention$a_m(w)$ の値が高くなっている単語をストップワードとすることによって，予測を誤る要因が削除され分類精度が改善すると考えられる．

正解データでの Attention$a_c(w)$ の値が高くなっている単語は，分類器が正しい予測をした際に注目していた単語であると解釈できる．不正解データでの Attention$a_m(w)$ の値が高くなっている単語であっても，正解データでの Attention$a_c(w)$ の値が高くなっている単語は分類をするうえで役立つ単語であり，ストップワードとしてデータから削除することが適切ではないと考えられる．しかし，不正解データでの Attention$a_m(w)$ の値に着目してストップワードを生成するという方法では，予測を誤る要因と同時に正しく予測するための重要な特徴を削除してしまう可能性がある．

正解データでの Attention$a_c(w)$ の値が小さい単語は，正しい予測をすることにあまり貢献していないと考えられる．不正解データでの Attention$a_m(w)$ の値が大きく，正解データでの Attention$a_c(w)$ の値が小さい単語をストップワードとするのが良いと考えた．不正解データでの Attention$a_m(w)$ の値と正解データでの Attention$a_c(w)$ の値の差に着目した．Attention の差 $a_d(w)$ は以下の式で求めた．

$$a_d(w) = a_m(w) - a_c(w)$$

Attention の差 $a_d(w)$ の値が大きい単語ほど分類に悪影響を与えていると考えられる．そのため，Attention の差 $a_d(w)$ の値が大きい単語をストップワードにするべきであると考えられる．

3.2 ストップワードの自動生成

Attention の差 $a_d(w)$ の値に着目してストップワードを生成する．ストップワードとなる単語を決める方法には，Attention の差 $a_d(w)$ の値に対して閾値を決める方法や，ストップワードとなる単語の数を決める方法が存在する．しかし，上記の手法は閾値やストップワードの数を手動で決める必要がある．本研究では，ストップワードの生成を自動で行う．そのために，Attention の差 $a_d(w)$ の値をストップワードとなる確率として用いた．Attention の差 $a_d(w)$ の値が大きい単語ほどストップワードになる確率が高くなるように設定した．

Attention の差 $a_d(w)$ の値が取る範囲はデータセットや分類の基準によって異なる．データセットや分類の基準ごとに $a_d(w)$ の値が最大のものを 1, 最小のものを 0 とするように正規化をした．Attention の差 $a_d(w)$ 全体の集合を A とする．Attention の差 $a_d(w)$ の最大値を $\max A$，最小値を $\min A$ とする．以下の式で正規化した Attention の差 $a_n(w)$ を求めた．

$$a_n(w) = \frac{a_d(w) - \min A}{\max A - \min A}$$

正規化した Attention の差 $a_n(w)$ の値をそれぞれの単語がストップワードとなる確率として用いる．単語 w がストップワードとなる確率 $p(w)$ は以下の式で表せる．

$$p(w) = a_n(w)$$

我々は，$p(w)$ を用いて 2 種類の方法でストップワードの生成をした．ストップワード生成の手順を以下に示す．

3.2.1 手法 1：ストップワードリスト

ストップワードリストを作成し，ストップワードリストに含まれる単語を入力文書中から削除する方法である．

システム開発者は，ストップワードリストに含まれる単語を入力文書中から削除するという方法でストップワードの除去を行うことがある．しかし，公開されている既存のストップワードリストは分類精度の向上に有効ではないことが分かっている．そのため，分類精度の向上に有効なストップワードリストを新たに作成する必要があると考えた．

$p(w)$ を単語 w がストップワードになる確率とする．

以下に示す手順でストップワードリストを作成する．

（1） w と $p(w)$ が格納されたリストに含まれる単語 w に対して，確率 $p(w)$ に従って w がストップワードになるかどうかの判定を行う．

（2） ストップワードになると判定された単語 w をストップワードリストに追加する．

$p(w_1)$ が 0.9 となる単語 w_1 が存在した場合，単語 w_1 は 9 割の確率でストップワードリストに追加される．一方で，$p(w_2)$ が 0.1 の単語 w_2 は，ストップワードリストに追加される確率が 1 割になる．つまり，$p(w)$ が大きい単語はストップワードリストに追加されやすい．

我々は，以下にストップワード除去を行う手順を示す．

（1） 入力文書中に出現する単語がストップワードリスト内に存在するか確認する．

（2） 単語 w がストップワードリストに含まれている場合，入力文書中に出現する単語 w を全て削除する．

3.2.2 手法 2：確率的ランダム除去

入力文書中に出現する単語一つ一つに対して削除するかどうかを確率的に決める方法である．手法 1 のストップワードリストに含まれる単語を全て削除する方法とは異なり．手法 2 では同じ単語であっても削除する場合と削除しない場合がある．

$p(w)$ が大きい単語が分類精度に悪影響を与えていると考えられる．しかし，$p(w)$ が大きい単語が必ずしも分類精度に悪影響を与えているとは言えない．$p(w)$ が大きい単語 w であっても，入力文書中に存在する w を全て削除するのではなく，一部を削除しないで入力文書中に残す方法を考える．手法 1 の問題点として $p(w)$ が大きい単語であってもストップワードリストに追加されず，分類に悪影響を与えていると考えられる単語が全く削除されない場合があると考えられる．ストップワードリストの作成がうまくいくかどうかに依存する．手法 1 はストップワードリストを自動で生成するものである．手法 2 ではストップワードリストの作成をしないで，ストップワードの除去を自動で行う．一つ一つの単語に対して削除するかの判定を行うことによって，$p(w)$ が大きい単語を多く削除し，$p(w)$ が小さい単語の削除する数を少なくすることができると考えた．

ストップワードリストの作成を行わずにストップワードの除去を行う．$p(w)$ を単語 w がストップワードとして削除される確率とする．以下にストップワード除去を行う手順を示す．

（1） 入力文書中の単語一つ一つに対して，ストップワードになるかどうか判定する．

（2） ストップワードになると判定された単語を入力文書中から削除する．

3.2.1 節に示したストップワードリストを作成する方法とは異なり，同じ単語であっても確率によって除去される場合と除去されない場合がある．$p(w_1)$ が 0.9 の単語 w_1 が存在した場合，w_1 は 9 割の確率で入力文書中から削除される．単語 w_1 が入力文書中に 100 個存在した場合，およそ 90 個の w_1 が削除されると思われる．一方で，$p(w_2)$ が 0.1 の単語 w_2 は 1 割の確率で入力文書中から削除される．単語 w_2 が入力文書中に 100 個存在した場合，およそ 10 個の w_2 が削除されると思われる．$p(w)$ が大きい単語ほど，削除されやすくなる．

4 実　　験

本実験では，提案手法によるストップワード除去の有効性を確かめた．ストップワードを除去したデータの分類を行った場合とストップワードを除去せずに分類を行った場合で分類精度を比較した．精度に有意な差があるかどうか統計的検定によって示した．

最適なストップワードはデータセットや分類の目的によって異なると考えられる．4.2 節に示すデータセットで 2 種類，4.3 節に示すデータセットと，4.4 節に示すデータセットでそれぞれ 1 種類ずつの，合計で 4 種類の分類を行った．

4.1 実 験 手 順

以下に実験の手順を示す．

（1） データセットごとに各ラベルのデータ数が均等になるようにデータを収集する．データ数を均等にする方法は，データセットごとに異なるため 4.2 節，4.3 節，4.4 節で説明する．

（2） 訓練用，検証用，テスト用のデータ数の比率が 8：1：1 になるように（1）で収集したデータを分割する．

（3） BERT を用いて訓練用データで学習を行う．BERT には，東北大学乾・鈴木研究室が公開している学習済みモデル [1] を使用した．最大エポック数を 10,000 エポックとした．検証用データの損失の最小値が 50 エポックの間更新されなかった場合に学習を停止し，検証用データの損失が最も小さいモデルを採用する．

（4） 3 章で示した手順でストップワードの除去を行う．テスト用データを入力とする．ストップワードリストを用いてストップワードの除去を行った場合と，確率的ランダム除去でストップワードの除去を行った場合で，それぞれ実験を行った．これら 2 種類の手法を同時に使用することはしない．

（5） ストップワードを除去した訓練用データを用いて BERT で学習を行う．学習の条件は（3）と同じにした．ストップワードの除去方法ごとにモデルを 10 個ずつ作成した．

（6） （5）で作成したモデルを用いてテスト用データの分類を行う．10 個のモデルでそれぞれテスト用データの分類を行い，精度の平均を比較する．

（7） 検定を行う．(6) で求めた精度の平均に有意な差があるかどうか確かめる．帰無仮説は「ストップワードの使用による精度の変化に有意な差はない.」である．有意水準は 5 %で，対応のない 2 標本 t 検定を行った．p 値が 0.05 を下回った場合に帰無仮説が棄却され，精度の変化に有意な差があるといえる．

4.2 楽天データセット

楽天データセット [2] は国立情報学研究所から公開されているデータセットである．楽天市場のデータは，商品データ約 2 億 8,000 万件，商品レビューデータ約 7,000 万件，ショップレビューデータ約 2,250 万件を含む．本研究では，楽天市場の商品レビューデータの，レビュー本文，評価点，使い道の項目を使用した．各ラベルのデータが 2,000 件になるようにデータ収集を行った．

4.2.1 実験 1．楽天データセット：評価点

a） データ内容

ポジティブ，ネガティブ，ニュートラルのラベルの作成にはデータセットの評価点の項目を使用した．評価点が 1，2 のデータをネガティブ，評価点が 3 のデータをニュートラル，評価点

1：https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

2：楽天グループ株式会社 (2020): 楽天市場データ．国立情報学研究所情報学研究データリポジトリ．(データセット). https://doi.org/10.32130/idr.2.1

表 1　実験結果：分類精度と t 検定を行った際の p 値

$p(w)$	手法	実験 1		実験 2		実験 3		実験 4	
		精度	p 値	精度	p 値	精度	p 値	精度	p 値
ストップワードなし		0.6828	-	0.5412	-	0.7297	-	0.8971	-
不正解データの Attention	リスト	0.6658	2.406×10^{-3}	0.5000	9.143×10^{-5}	0.6738	7.929×10^{-6}	0.8971	9.999×10^{-1}
	ランダム	0.6648	3.310×10^{-3}	0.4883	2.901×10^{-7}	0.6983	3.372×10^{-3}	0.8911	7.238×10^{-2}
Attention の差	リスト	0.6858	5.270×10^{-1}	0.5413	9.905×10^{-1}	0.7364	4.587×10^{-1}	0.8969	9.468×10^{-1}
	ランダム	0.6838	8.394×10^{-1}	0.5475	4.170×10^{-1}	0.7307	9.126×10^{-1}	0.9013	2.343×10^{-1}
正規化した Attention の差	リスト	0.6923	1.157×10^{-1}	0.5395	7.686×10^{-1}	0.7360	5.356×10^{-1}	0.9016	2.583×10^{-1}
	ランダム	0.6835	8.959×10^{-1}	0.5476	2.946×10^{-1}	0.7326	7.646×10^{-1}	0.8984	7.471×10^{-1}

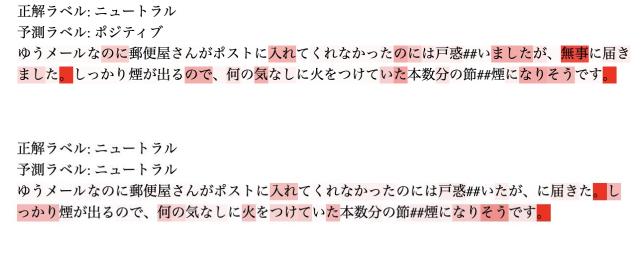

図 1　実験 1 でストップワードの除去を行う前後での Attention の比較

が 4，5 のデータをポジティブとして，3 クラスの分類を行った．

b）　結果・考察

表 1 の実験 1 の列にテスト用データ分類時の分類精度と t 検定を行った際の p 値を示す．$p(w)$ の列にはストップワードになる確率として用いた値を示しており，「なし」がストップワードの除去を行わずに分類した場合である．手法の列にはストップワード除去の方法を表している．リストは 3.2.1 節に示したストップワードリストを用いる場合を示しており，ランダムは 3.2.2 節に示した確率的ランダム除去を行う場合を示している．精度を比較すると，正規化した Attention の差に着目してストップワードリストを用いてストップワードの除去を行った場合に最も精度が高いことが分かる．

不正解データの Attention に着目してストップワードの除去を行った場合にはどちらの手法についてもストップワードの除去をしない場合よりも精度は低く，t 検定を行った際の p 値が 0.05 を下回っているため精度の低下に有意な差が認められた．そのため，不正解データの Attention に着目するよりも，Attention の差や正規化した Attention の差に着目してストップワードを生成した方が良いといえる．

不正解データの Attention，Attention の差，正規化した Attention の差のどれを $p(w)$ として用いた場合にも，確率的ランダム除去よりもストップワードリストの方が精度が高かった．楽天データセットを用いたポジティブ，ネガティブ，ニュートラル分類の場合にはストップワードリストを用いてストップワードの除去を行う方が良いといえる．

図 1 に正規化した Attention の差に着目し，ストップワードリストを用いてストップワードの除去を行ったデータの例を示す．分類を行った際の Attention が高い単語ほど色が濃くなっている．色の濃い単語は，分類器が分類の際に注目した単語であると考えられる．ストップワードの除去を行う前には正解ラベルと異なるラベルが予測されていることが分かる．ストップワードの除去を行う前には「無事」という単語の Attention が高い．ストップワードリストを用いてストップワードの除去を行うことによって，「無事」という単語がストップワードとして入力文書中から除去された．ストップワードの除去を行った後のデータは正解ラベルと同じラベルが予測された．予測を誤る要因となる単語をうまく除去できているといえる．

Attention の差の最大値は 0.52 であった．Attention の差を用いてストップワードの除去を行った場合，ストップワードとなる確率が最大の単語でもおよそ半分の確率でしか除去されない．正規化を行うことによって多くの単語で $p(w)$ が大きくなった．その結果ストップワードとして除去される単語が増加した．

ストップワードリストに含まれる単語の数は，不正解データの Attention を用いた場合には平均で 142 個，Attention の差を用いた場合には平均で 38 個，正規化した Attention の差を用いた場合には平均で 77 個であることを確認した．不正解データの Attention を用いた場合にストップワードが多く設定される．そのため，分類に有益な単語もストップワードになってしまっていることが考えられる．正規化を行うことによって，正規化を行う前と比較して，ストップワードリストに含まれる単語の数はおよそ 2 倍に増加した．また，正規化した Attention の差を用いた場合の方が分類精度が高かった．Attention の差について正規化をすることによって，予測を誤る要因となる単語をより多く除去できるようになったといえる．

Attention の差に着目したストップワードリストを用いた場合でも，ストップワードを使用せずに分類を行った場合に比べて精度が高いため，予測を誤る要因となる単語を除去できているといえる．正規化した Attention の差に着目したストップワードリストを用いた場合には，Attention の差に着目したストップワードリストを用いた場合に比べてストップワードリストに含まれる単語の数が多いことに加えて，テスト用データ分類時の精度も高いことが分かる．Attention の差に着目したストップワードリストを用いた場合に比べて，予測を誤る要因となる単語を多く除去できているといえる．一方で，確率的ランダム除去を用いた場合には，ストップワードリストを用いた場合に比べて精度が向上しなかったことが分かる．予測を誤る要

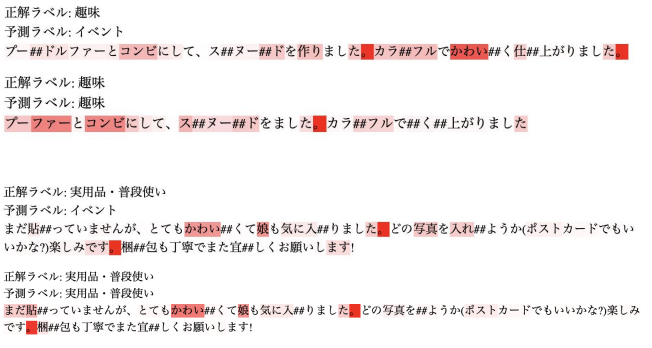

図 2　実験 2 でストップワードの除去を行う前後での Attention の比較

因となる単語を十分に除去できなかったと考えられる．

楽天データセットを用いたポジティブ，ネガティブ，ニュートラル分類の場合には，Attention の差や正規化した Attention の差に着目してストップワードリストを用いてストップワードの除去を行う方が良いといえる．

4.2.2　実験 2．楽天データセット：使い道

a）　データ内容

データセットの使い道の項目をラベルとして使用した．使い道の項目には,「実用品・普段使い」,「プレゼント」など 6 種類のラベルが付与されている．付与されたラベルを用いて 6 クラスの分類を行った．

b）　結果・考察

表 1 の実験 2 の列にテスト用データ分類時の分類精度と t 検定を行った際の p 値を示す．精度を比較すると，正規化した Attention の差の値に着目して確率的ランダム除去を行った場合に最も精度が高いことが分かる．

不正解データの Attention のみに着目してストップワードの除去を行った場合には，どちらの手法についてもストップワードの除去をしない場合よりも精度は低く，t 検定を行った際の p 値が 0.05 を下回っているため精度の低下に有意な差が認められたことが確認できる．

Attention の差に着目した場合も正規化した Attention の差に着目した場合も，ストップワードリストを用いるよりも確率的ランダム除去の方が精度が高いことが分かる．楽天データセットを用いた商品の使い道の分類には，確率的ランダム除去でストップワードの除去を行った方が良いといえる．

確率的ランダム除去の精度に注目すると Attention の差の値について正規化を行う前後での精度の変化が小さいことが分かる．正規化を行う前の Attention の差の最小値が 0 で，最大値が 0.92 であったため，値の範囲が 0 から 1 になるように正規化を行ってもあまり値が変化しなかった．そのため，それぞれの単語の除去される確率があまり変化せず，精度にもあまり差が出なかったと考えられる．

図 2 に，正規化した Attention の差に着目して，確率的ランダム除去でストップワードの除去を行った際の例を二つ示す．上に示す例では,「かわい」という単語がストップワードの除去を行う前に Attention が高く，誤った予測をする要因になっていると考えられる．確率的ランダム除去でストップワードの除去を行うことによって,「かわい」という単語がストップワードとして入力文書中から除去された．予測を誤る要因となる単語が削除されたことによって予測がうまくいくようになったと考えられる．下に示す例では上に示した例とは異なり，ストップワードの除去を行っても,「かわい」という単語が削除されなかった．図 2 から，ストップワードの除去を行った後には，ストップワードの除去を行う前と比較して,「かわい」という単語の Attention が高くなっていることがわかる.「かわい」という単語に注目することによって，正しく予測できるようになったと考えられる.「かわい」という単語を削除することによってうまく分類できるようになったと考えられるデータと,「かわい」という単語を削除しないことによってうまく分類できるようになったと考えられるデータが存在した．誤った予測をした際に，Attention が高くなっている単語であっても，必ずしも分類性能に悪影響を与えているとは言えず，全て削除してしまうことが適切では無いといえる．

楽天データセットを用いた商品の使い道での分類には，正規化した Attention の差の値に着目して確率的ランダム除去でストップワードの除去を行うと良いといえる．

4.3　実験 3．Twitter 日本語評判分析データセット

a）　データ内容

Twitter 日本語評判分析データセット [10][3]は，岐阜大学鈴木研究室が公開しているデータセットである．データセットには，2015 年から 2016 年頃のツイートのツイート ID と携帯電話などのジャンルを表す ID，ツイート本文を取得するための status ID とポジティブ，ネガティブ，ニュートラルのラベルが含まれている．実験ではツイートの本文とポジティブ，ネガティブ，ニュートラルのラベルを使用した．データセットの中には，複数のラベルが付与されたデータが存在する．ポジティブ，ネガティブ，ニュートラルのいずれか一つのラベルが付与されたデータを使用して 3 クラスの分類を行った．各ラベルのデータが 1,400 件になるようにデータの収集を行った．

b）　結果・考察

表 1 の実験 3 の列にテスト用データ分類時の分類精度と t 検定を行った際の p 値を示す．精度を比較すると，Attention の差に着目してストップワードリストを用いてストップワードの除去を行った場合に最も精度が高いことが分かる．

不正解データの Attention に着目してストップワードの除去を行った場合にはどちらの手法についてもストップワードの除去をしない場合よりも精度は低く，t 検定を行った際の p 値が 0.05 を下回っているため精度の低下に有意な差が認められた．そのため，不正解データの Attention に着目するよりも，Attention の差や正規化した Attention の差に着目してストップワードを生成した方が良いといえる．

Attention の差に着目した場合も正規化した Attention の差に着目した場合も，ストップワードリストの方が確率的ランダム除去よりも精度が高かった．Twitter データを用いてポジティ

3：https://www.db.info.gifu-u.ac.jp/sentiment_analysis/

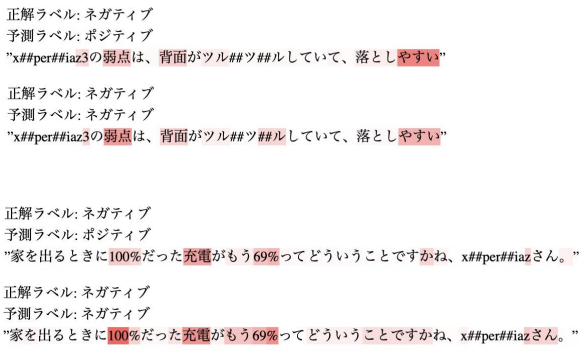

図 3　実験 3 でストップワードの除去を行う前後での Attention の比較

ブネガティブニュートラル分類を行う場合には，ストップワードリストを用いてストップワード除去を行う方が良いといえる．

ストップワードリストに含まれる単語の数は，不正解データの Attention を用いた場合には平均で 493 個，Attention の差を用いた場合には平均で 21 個，正規化した Attention の差を用いた場合には平均で 50 個であった．Attention の差の値について正規化を行うことによって，ストップワードリストに含まれる単語の数はおよそ 2 倍に増加した．正規化を行う前の Attention の差を用いた場合の方が分類精度が高かった．正規化を行うことによって，分類に重要な特徴を持った単語がストップワードになりやすくなったと考えられる．

図 3 にストップワードの除去によって正しいラベルを予測できるようになったデータの例を示す．どちらの例も，ストップワードの除去を行わずに分類した際の Attention を可視化したものを上に示す．下には Attention の差に着目してストップワードリストを用いてストップワードの除去を行って分類した場合の Attention を可視化したものを示す．図 3 に示した二つの例について，どちらの場合にもストップワードの除去を行う前後で入力文書に変化はない．これらのデータについては，予測を誤る要因となっていそうな単語が除去されていないといえる．しかし，ストップワードの除去を行う前後で予測ラベルが変化していることが分かる．Attention が変化したことによって正しく分類できるようになったと思われる．ストップワードの除去を行ったデータを用いて学習させることによって，Attention が変化し正しく予測できる場合もある．ストップワードの除去を行うことによって，ストップワードとして除去される単語を含まないデータにも影響を与えていることが分かる．

Twitter データを用いてポジティブ，ネガティブ，ニュートラル分類では，Attention の差に着目してストップワードリストを用いてストップワードの除去を行うと良いといえる．

4.4　実験 4．livedoor ニュースコーパス

a）　データ内容

livedoor ニュースコーパス [4]は，株式会社ロンウィットが公開しているデータセットで，NHN Japan 株式会社が運営する livedoor ニュースのデータを収集したものである．データセットにはニュース記事の URL，日付，タイトル，本文を含むデータが 7,376 件存在する．ニュース記事は九つのカテゴリに分かれている．各カテゴリには 512 から 901 件のデータが存在する．ニュース記事の本文からカテゴリを予測する 9 クラスの分類を行った．このデータセットでは，カテゴリごとのデータ数に偏りがあった．そのため，データ数の少ないカテゴリに合わせて，各ラベルのデータが 500 件になるようにデータを収集した．

b）　結果・考察

表 1 の実験 4 の列にテスト用データ分類時の分類精度と t 検定を行った際の p 値を示す．精度を比較すると，正規化した Attention の差に着目してストップワードリストを用いてストップワードの除去を行った場合に最も精度が高いことが分かる．

不正解データの Attention に着目したストップワードを用いた場合，ストップワードリストでは精度に変化がなく，確率的ランダム除去では精度が下がっている．そのため，不正解データの Attention のみに着目したストップワードは有効とは言えない．Attention の差や正規化した Attention の差に着目してストップワードの生成を行う方が良いといえる．

ストップワードリストの精度に注目すると，正規化した Attention の差を用いた場合の方が精度が高いことが確認できる．ストップワードリストに含まれる単語の数は，不正解データの Attention を用いた場合には平均で 116 個，Attention の差を用いた場合には平均で 53 個，正規化した Attention の差を用いた場合には平均で 65 個であった．Attention の差を用いた場合よりも正規化した Attention の差を用いた場合の方がストップワードリストに含まれる単語の数は多い．正規化した Attention の差を用いることによって，ストップワードになりやすい単語をうまく設定できたと考えられる．Attention の差を用いた場合には，ストップワードを使用しない場合よりも精度が低い．Attention の差を用いた場合には，予測を誤る要因となる単語を十分に除去できていなかったと考えられる．ストップワードとして除去する単語の数が少ない場合に分類精度が下がる可能性があると考えられる．

livedoor ニュースコーパスを用いたニュースカテゴリの予測では，正規化した Attention の差に着目してストップワードリストを用いてストップワードの除去を行うと良いといえる．

5　おわりに

本研究では，BERT を用いた文書分類タスクの精度向上を目的として，システムが自動でストップワードを生成する手法を提案した．我々は，BERT を用いて文書分類を行った際の入力文書中に出現する単語の Attention に着目して，ストップワードの生成を行うシステムを構築した．

不正解データに出現した場合に Attention の高い単語は，分類器が誤った予測をする要因になっていると考えられる．そのため，不正解データの Attention に着目し，ストップワードの生成を行うと良いと考えた．

しかし，不正解データの Attention のみに着目した場合，正解データと不正解データのどちらに出現した場合にも Attention

4：http://www.rondhuit.com/download.html\#ldcc

の高い単語がストップワードになることが考えられる．正解データにおいて Attention が高い単語は，分類器が正しい予測をすることに貢献している可能性があると考えられる．そのため，不正解データの Attention のみに着目してストップワードの生成を行うことは適切ではないと考えた．

正解データで Attention の低い単語は，分類器が正しい予測をすることに貢献していないと考えられる．そのため，不正解データにおいて Attention が高く，正解データにおいて Attention の低い単語をストップワードとすると良いと考えた．そこで，単語が不正解データに出現した場合と正解データに出現した場合との Attention の差に着目した．

ストップワードの生成方法を 2 種類提案した．一つ目は，単語が不正解データに出現した場合と正解データに出現した場合との Attention の差をストップワードリストに含まれる確率として扱い，ストップワードリストを作成する手法である．作成したストップワードリストに含まれる単語を入力文書中から全て削除する．二つ目は，単語が不正解データに出現した場合と正解データに出現した場合との Attention の差を単語が除去される確率として扱う手法である．入力文書に出現する単語一つ一つについて，確率を用いてストップワードとして入力文書中から削除するかどうかの判定を行う．

提案手法を用いて生成したストップワードの有効性を確かめるために実験を行った．自動で生成したストップワードを用いて BERT で文書分類を行った．そして，ストップワードを使用せずに分類を行った場合と精度の比較を行った．さらに，ストップワードの除去による精度の変化に有意な差があるかどうか確認するために統計的検定を行った．

実験の結果，ストップワードを使用することによって，ストップワードを使用せずに分類を行った場合よりも高い精度で分類できることが分かった．不正解データにおける Attention の値と正解データにおける Attention の値との差が大きい単語を，ストップワードとして除去したことによって，正しいラベルが予測されたデータの存在を確認できた．不正解データにおいて Attention の高い単語は予測を誤る要因になっており，入力文書から削除することによって分類精度が向上したと考えられる．ストップワードの除去を行ったデータを用いて学習することによって，分類時の Attention が変化したことが確認できた．

今回行った四つの実験のうち三つの実験では，不正解データにおける Attention のみに着目して生成したストップワードを用いた場合には，精度の低下に優位な差が認められたことを確認した．不正解データにおける Attention のみに着目すると，正しい予測を行う上で重要な特徴も削除されてしまったと考えられる．そのため，不正解データにおいて Attention の高い単語が，必ずしも予測を誤る要因になっているとは言えない．不正解データにおける Attention の値のみではなく，正解データにおける Attention の値との差に着目することによって，分類精度を向上させるのに有効なストップワードを生成することができると考えられる．

不正解データにおける Attention の値と正解データにおける Attention の値の差について正規化を行うことによって，多くの単語でストップワードとなる確率が上昇することを確認した．ストップワードとなる確率が上昇したことによって，ストップワードとなる単語の数が増加したことも確認した．実験結果から，実験 1，実験 2，実験 4 においては Attention の差を正規化した値に着目したストップワード生成を行った場合に最も分類精度が高くなっていることが分かる．不正解データにおける Attention の値と正解データにおける Attention の値の差について正規化を行うことによって，正規化を行う前に比べて分類精度の向上に適したストップワードが得られるといえる．

今回生成したストップワードによって分類精度の改善が見られる場合もあった．Attention に着目することによって，有意な差はないが精度の改善は見られた．今後は，BERT を用いた文書分類タスクにおいて，有意に精度を改善するストップワードを作成したいと考えている．

謝辞　本研究の一部は JSPS 科研費 19H04218 および越山科学技術振興財団の助成を受けたものです．本研究では、国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（`https://rit.rakuten.com/data_release/`）を利用しました．

文　献

[1] Rajaraman Anand and Ullman Jeffrey David. *Mining of massive datasets*. Cambridge university press, 2011.

[2] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 810–817, 2014.

[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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] 桑原悠希, 鈴木優. BERT を用いた文書分類タスクにおけるストップワードの有効性の検証. 研究報告情報基礎とアクセス技術 (IFAT), Vol. 2022, No. 41, pp. 1–6, 2022.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[6] 國府久嗣, 山崎治子, 野坂政司. 内容推測に適したキーワード抽出のための日本語ストップワード. 日本感性工学会論文誌, Vol. 12, No. 4, pp. 511–518, 2013.

[7] Saziyabegum Saiyed and Priti Sajja. Empirical analysis of static and dynamic stopword generation approaches. In *ICT Systems and Sustainability*, pp. 149–156. Springer, 2022.

[8] 木村優介, 駒水孝裕, 波多野賢治. ストップフレーズ抽出を併用した文書分類. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022), 2022. A23-4.

[9] Christopher Fox. A stop list for general text. In *Acm sigir forum*, Vol. 24, pp. 19–21. ACM New York, NY, USA, 1989.

[10] Yu Suzuki. Filtering method for twitter streaming data using human-in-the-loop machine learning. *Journal of Information Processing*, Vol. 27, pp. 404–410, 2019.