卒業論文

Attention に着目した確率的ストップワード生成手法

桑原 悠希

2023 年 2 月 8 日

岐阜大学 工学部 電気電子・情報工学科 情報コース
鈴木研究室

本論文は岐阜大学工学部に
学士（工学）授与の要件として提出した卒業論文である.

桑原 悠希

指導教員：
鈴木 優 准教授

Attention に着目した確率的ストップワード生成手法*

桑原 悠希

内容梗概

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．研究者やシステム開発者は，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で文書分類の精度が向上すると考えられる．適切なストップワードはデータセットや分類の基準によって異なると考えられる．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．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ことで分類精度を改善できると考えた．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可能性があると考えられる．そのため，不正解データに出現した単語の 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．正解データに出現した場合の Attention の値が小さい単語は，分類器が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．そのため，不正解データに出現した場合の Attention の値が大きく，正解データに出現した場合の Attention 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．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 の差に着目する．ストップ

*岐阜大学 工学部 電気電子・情報工学科 情報コース 卒業論文, 学籍番号: 1193033061, 2023 年 2 月 8 日.

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，ストップワードの生成を行う場合にはどちらか片方の手法のみで行う．一つ目の手法では，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．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．二つ目の手法では，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．入力文書に出現する単語一つ一つについて，確率を用いて入力文書中から削除するかどうかの判定を行う．自動で生成したストップワードが分類精度の向上に有効かどうかを確認するために実験を行った．BERT を用いて，ストップワードを除去したテキストデータの分類を行い，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．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．分類を行った際の Attention の値に着目することによって，分類精度を向上させることのあるストップワードが生成できることが分かった．また，不正解データに出現した単語の Attention のみに着目するよりも，単語が正解データに出現した場合との Attention の差に着目した方が，分類精度が高くなることが分かった．

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

目次

図目次

表目次

第 1 章　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]．研究者やシステム開発者は，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を向上させることに有効ではないことが分かっている [4]．そのため，分類精度を向上させること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，ストップワードの生成を行うことによって，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可能性があると考えられる．そのため，単語が不正解データに出現した場合の 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．正解データに出現した場合の Attention の値が小さい単語は，分類器が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．不正解データに出現した場合の Attention の値が大きく，正解データに出現した場合の Attention 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．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．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対して，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を確率的に決める．

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が有意なもの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やストップワードの効果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．とくに，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．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くなる場合もあった．

本研究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．

本論文の構成は以下の通りである．2 章では本論文で用いた技術，及び手法について述べる．3 章では本研究と関連のある先行研究について述べる．4 章では本論文で提案するストップワードの生成手法について述べる．5 章では 4 章で述べた提案手法を用いて生成したストップワードが有効に機能するか検証するために行った実験の内容と結果，考察について述べる．最後に 6 章では本論文のまとめと今後の課題について述べる．

第 2 章　基本的事項

本論文で用いた技術，及び手法について述べる．

2.1　ストップワード

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．研究者やシステム開発者は，検索や分類の精度を向上させるためにストップワードを用いることがある．このとき，ストップワードリストとして公開されているものを使用する，テキストデータ中の単語の出現回数に着目する，品詞情報を使用するなどの方法でストップワードを決める．

ストップワードリストとして公開されているものには，Slothlib[6] や GiNZA[7] のストップワードリストが存在する．Slothlib のストップワードリスト*には「あそこ」や「年」,「上」や漢数字など 310 個の単語が存在する．GiNZA のストップワードリストには「あまり」や「きっかけ」,「た」や「らしい」など 154 個の単語が存在する．

2.2　BERT

BERT とは，Bidirectional Encoder Representations from Transformers の略称である．2018 年に Google の Jacob Devlin らによって発表された自然言語処理モデルである．BERT は，Transformer[8] の Encoder を使用したモデルである．

BERT には，文脈を考慮した分散表現を獲得できるという特徴がある．同じ単語であっても文脈によって異なる分散表現が得られる．

事前学習では，Masked Language Model（MLM）と Next Sentence Prediction（NSP）の 2 種類のタスクが行われている．事前学習を行う際にはラベルが付与されていないデータを用いる．MLM は，別の単語に置き換えられた単語を周辺の単語から予測するというタスクである．入力文の 15 %を別の単語に置き換える．そ

*http://svn.sourceforge.jp/svnroot/Slothlib/CSharp/Version1/Slothlib/NLP/Filter/StopWord/word/Japanese.txt

のうち，80 %は [MASK] トークンに，10 %はランダムな別の単語に置き換えられ，10 %は置き換えられずに元の単語のままになる．入力文の置き換えられた部分について，置き換えられる前の単語の予測を行う．MLM を行うことによって，モデルは単語に対応する文脈情報を学習する．NSP は，二つの文のペアを入力とし，文の関係性を予測するタスクである．二つの入力文のうち，片方を 50 %の確率でランダムな別の文に置き換える．入力された二つの文が連続した文かどうかの予測を行う．NSP を行うことによって，モデルは文単位での意味表現を獲得することができる．

ファインチューニングとは，学習済みのモデルを特定のタスクを解くことに特化するように再学習させることである．BERT では，事前学習によって獲得したモデルの重みを初期値としてファインチューニングを行う．ファインチューニングを行う際には，ラベル付きのデータを使用する．

2.3 Attention

Attention 機構 [5] 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．入力されたデータに対して，どの部分が重要であるか重み付けをする．そして，重要性を考慮したベクトル量として出力をする．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元々は自然言語処理の分野を中心に発展した技術であるが，現在は画像処理の分野でも利用されている．

2.2 節で説明した BERT は Attention 機構を用いたモデルである．BERT には Multi-Head Attention が用いられている．Multi-Head Attention は複数の Self-Attention モデルが並列になっている構造である．Self-Attention とは，ある 1 文に含まれている単語のみで計算された，単語間の関連度のようなものである．複数の Attention 機構を用いることによって，一つの単語に対して複数の異なる特徴を獲得することができる．

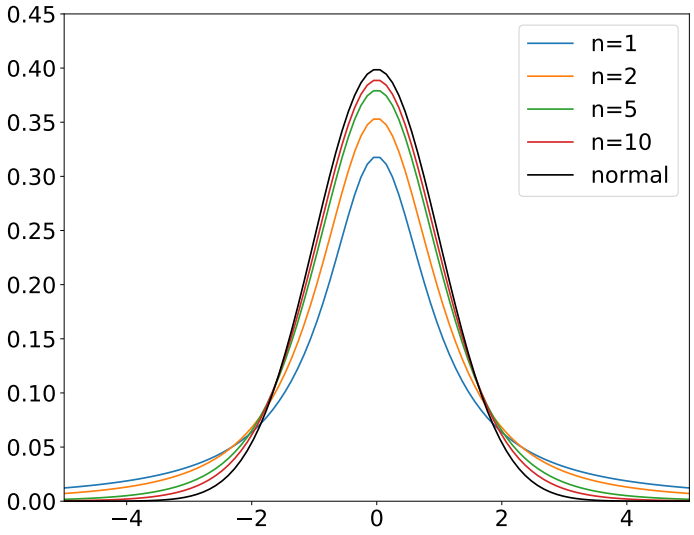

図 2.1　自由度 n を変化させたときの t 分布と正規分布

2.4　対応のない 2 標本 t 検定

対応のない 2 標本 t 検定とは，二つの独立した集団の間で平均に差があるかどうかを確かめたいときに用いられる手法である．t 分布は，標準正規分布に似た分布で，自由度というパラメータを変化させることによって分布の形が変わるという特徴を持っている．図 2.1 に示すように，自由度 n を大きくすることによって，t 分布は図中に normal で示されている標準正規分布に近づく．

本研究では，ストップワードの除去をしていないデータで学習したモデルとストップワードの除去をしたデータで学習したモデルとの分類精度の差に有意差があるか確認するために検定を行った．異なるモデルを使用しているため，対応のない 2 標本 t 検定を行った．

以下の手順で対応のない 2 標本 t 検定を行う．

1. 帰無仮説と対立仮説を設定する．

   対応のない 2 標本 t 検定では以下のように帰無仮説と対立仮説を設定する．

   - 帰無仮説：二つの標本の平均に有意な差はない．

- 対立仮説：二つの標本の平均に有意な差がある．

2. 有意水準を決める
   有意水準は帰無仮説を棄却するかどうかの基準である．本研究では，有意水準は 5 %とした．
3. t 値を求める．
   二つの標本の平均をそれぞれ，$\bar{x_1}$，$\bar{x_2}$ とする．二つの標本の母平均をそれぞれ，μ_1，μ_2 とする．二つの標本のデータ数をそれぞれ，n_1，n_2 とする．以下の式で表される検定統計量 t を求める．

$$t = \frac{(\bar{x_1} - \mu_1) - (\bar{x_2} - \mu_2)}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$= \frac{\bar{x_1} - \bar{x_2} - (\mu_1 - \mu_2)}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

   帰無仮説の「二つの標本の母平均が等しい.」が正しいと仮定すると, $\mu_1 - \mu_2 = 0$ となるため，検定統計量 t は以下のように表せる．

$$t = \frac{\bar{x_1} - \bar{x_2}}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

   このとき s は，以下に示す式で求められる s^2 の平方根を求めたものである．二つの標本の不偏分散を ${s_1}^2$，${s_2}^2$ とする．

$$s^2 = \frac{(n_1 - 1){s_1}^2 - (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$$

4. p 値を求める．
   t 分布をもとに，p 値を求める．p 値は，帰無仮説が成立する確率のようなものである．
5. p 値と有意水準を比較する．
   4 で求めた p 値と 2 で設定した有意水準を比較する．p 値が有意水準を下回った時に帰無仮説が棄却される．帰無仮説が棄却されることによって，対立仮説が採択され，二つの標本の平均に有意な差があると認められる．

2.5　評価指標

本研究ではモデルの性能を評価するための評価指標に Accuracy を用いた．Accuracy とは，正解率のことである．Accuracy は，モデルによって出力された予測ラベルがどの程度正解ラベルと一致していたかを表す指標である．正解ラベルと同じラベルが予測されたデータの数を全データの数で割ることによって Accuracy を求めることができる．

表 2.1 に 2 値分類を行った際の混同行列を示す．正解ラベルが 1 で予測ラベルが 1 のデータの数を TP，正解ラベルが 0 で予測ラベルが 1 のデータの数を FP，正解ラベルが 0 で予測ラベルが 0 のデータの数を TN，正解ラベルが 1 で予測ラベルが 0 のデータの数を FN とする．Accuracy は以下の式で求められる．

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

分類を行う際の各クラスのデータ数に偏りがある場合には，Accuracy が有効な評価指標とは言えない可能性がある．例えば，100 個のデータが存在し，そのうち 90 個の正解ラベルが 0 であったとする．このとき，モデルが全てのデータを 0 と予測すれば，Accuracy は 0.9 になる．Accuracy は高くなるがラベル 1 を全く予測しないため，有用なモデルであるとは言えなさそうである．本研究では，実験を行う際に各クラスのデータ数が同じになるようにデータの収集を行った．データ数に偏りがないため Accuracy での評価をした．

表 2.1　2 値分類の際の混同行列

		予測ラベル	
		0	1
正解ラベル	0	TN （True Negative）	FP （False Positive）
	1	FN （False Negative）	TP （True Positive）

第 3 章　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 [9] 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らか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 [10] 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．静的手法では，44.53 %，動的手法で 52.53 %も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と確認された．

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 [11] 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．単語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 [12]．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で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をストップフレーズとした．先行研究では単語の出現頻度に着目している．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の生成を行った．

第 4 章　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，テキストデータ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．
   - 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）ストップワードに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，テキストデータ中から全て削除する．
   - 確率的ランダム除去
     （a）入力のテキストデータ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テキストデータから削除する．

4.1 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，我々は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．

テキストデータ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 と正解データに出現した場合の Attention にそれぞれ着目し, 単語ごとに Attention 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_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 を $a(d_i, w_j)$ とする．不正解データに出現した場合の単語 w の Attention 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 の平均 $a_c(w)$ も求めた．

$a_m(w)$ 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することができる．$a_m(w)$ 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，予測を誤る要因となる単語が削除され分類精度が改善すると考えられる．

$a_c(w)$ 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することができる．$a_m(w)$ の値が高くなっている単語であっても，$a_c(w)$ 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の可能性がある．そのため，ストップワードとしてデータから除去することが適切ではないと考えられる．つまり，$a_m(w)$ の値のみ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．

$a_c(w)$ 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．$a_m(w)$ の値が大きく，$a_c(w)$ 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．そのため，$a_m(w)$ の値と $a_c(w)$ の値の差に着目した．Attention の差 $a_d(w)$ は以下の式で求めた．

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

$a_d(w)$ の値が大きい単語ほど分類に悪い影響を与えていると思われ，$a_d(w)$ 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．

4.2 ストップワードの自動生成

$a_d(w)$ 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，$a_d(w)$ の値で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，$a_d(w)$ の値を単語 w がストップワードとなる確率として用いた．$a_d(w)$ の値が大きい単語 w 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．

$a_d(w)$ の値が取る範囲はデータセットや分類の目的によって異なる．データセットや分類の目的ごとに $a_d(w)$ 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．$a_d(w)$ 全体の集合を A とする．$a_d(w)$ の最大値を $\max A$，$a_d(w)$ の最小値を $\min A$ とする．以下の式で正規化した $a_d(w)$ の値 $a_n(w)$ を求めた．

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

$a_n(w)$ 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．

4.2.1　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テキストデータから削除する方法である．研究者やシステム開発者は，ストップワードリストに含まれる単語をテキストデータ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を向上させるのに有効ではないことが分かっている．そのため，分類精度を向上させるの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 がストップワードリストに含まれる確率とする．以下に示す手順でストップワードリストを作成する．

1. 単語と Attention の差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 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 となる単語 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．

以下にストップワード除去を行う手順を示す．

1. テキストデータ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，テキストデータ中に出現する w を全てテキストデータ中から削除する．

4.2.2　手法 2：確率的ランダム除去

テキストデータ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可能性があると考えられる．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，テキストデータ中に存在する w を全て削除するのではなく，一部を削除しないでテキストデータ中に残す方法を考える．

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を作成せず，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どう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．

手法 2 では，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除去される確率とする．以下にストップワード除去を行う手順を示す．

1. テキストデータ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テキストデータから削除する．

4.2.1 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 割の確率でテキストデータから削除される．単語 w_1 がテキストデータ中に 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 割の確率でテキストデータから削除される．単語 w_2 がテキストデータ中に 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．

第 5 章　評価実験

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．提案手法による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使用せずに分類を行った場合で分類精度を比較した．Attention の差に着目することの有用性を確かめるために，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と分類精度を比較した．精度に有意な差がある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．5.2 節に示すデータセットで 4 種類，5.3 節に示すデータセットと，5.4 節に示すデータセットでそれぞれ 1 種類ずつの，合計で 6 種類の分類を行った．

5.1　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 5.2 節，5.3 節，5.4 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*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. 4 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，入力のテキストデータとする．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

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

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．

5.2 楽天データセット

楽天データセット[†]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，目的の項目を使用した．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．

5.2.1 実験 1．楽天データセット：評価点 2 クラス分類

データ内容

ポジティブネガティブ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．評価点は，1 から 5 の 5 段階である．評価点が 1，2 のデータをネガティブ，評価点が 4，5 のデータをポジティブとして，2 クラスの分類を行った．実験 1 において評価点が 3 のデータは使用しなかった．

[†]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

結果・考察

表 5.1 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．$p(w)$ の行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．手法の行はストップワード除去の方法を表している．ストップワードリストは 4.2.1 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，確率的ランダム除去は 4.2.2 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，ストップワードの除去を行わない場合に最も精度が高かった．ストップワードの除去を行ったものの中では，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った場合の精度が最も高いことが分かる．

t 検定を行った際の p 値を確認すると，不正解データの Attention に着目してストップワードの生成を行った場合と正規化した Attention の差に着目してストップワードの生成を行った場合に p 値が 0.05 を下回っていることが分かる．どちらの場合にもストップワードを使用せずに分類を行った場合と比べて精度が低い．楽天データセットを用いたポジティブネガティブ分類においては，不正解データの Attention に着目したストップワードと正規化した Attention の差に着目したストップワードは分類精度に悪影響を与えてしまうといえる．

Attention の差の値を正規化しても精度は改善されなかった．Attention の差の最大値は 0.87 であった．正規化を行うことによって多くの単語で $p(w)$ が大きく

表 5.1　実験 1 の分類精度と t 検定を行った際の p 値

$p(w)$	手法	精度	p 値
なし	-	0.9333	-
不正解データの Attention	ストップワードリスト	0.8985	1.367×10^{-5}
	確率的ランダム除去	0.8898	1.910×10^{-8}
Attention の差	ストップワードリスト	0.9303	4.111×10^{-1}
	確率的ランダム除去	0.9265	1.173×10^{-1}
正規化した Attention の差	ストップワードリスト	0.9200	8.341×10^{-3}
	確率的ランダム除去	0.9243	4.336×10^{-2}

なることが確認できた．正規化を行うことによって，分類に重要な特徴を削除する確率が上がってしまったことが考えられる．$a_c(w)$ の値よりも不正解データの Attention の値の方が大きい単語であっても，必ずしも分類精度に悪影響を与えているわけではなく，分類を行う上で重要な特徴になっている可能性があるといえる．

楽天データセットを用いたポジティブネガティブ分類において，今回提案したストップワードの生成方法は分類精度の向上に有効とは言えなかった．不正解データの Attention に着目してストップワードリストを使用してストップワードの除去を行った場合に精度の低下が最も小さかった．

5.2.2 実験 2．楽天データセット：評価点 3 クラス分類

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．

結果・考察

表 5.2 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

表 5.2　実験 2 の分類精度と t 検定を行った際の p 値

$p(w)$	手法	精度	p 値
なし	-	0.6828	-
不正解データの Attention	ストップワードリスト	0.6658	2.406×10^{-3}
	確率的ランダム除去	0.6648	3.310×10^{-3}
Attention の差	ストップワードリスト	0.6858	5.270×10^{-1}
	確率的ランダム除去	0.6838	8.394×10^{-1}
正規化した Attention の差	ストップワードリスト	0.6923	1.157×10^{-1}
	確率的ランダム除去	0.6835	8.959×10^{-1}

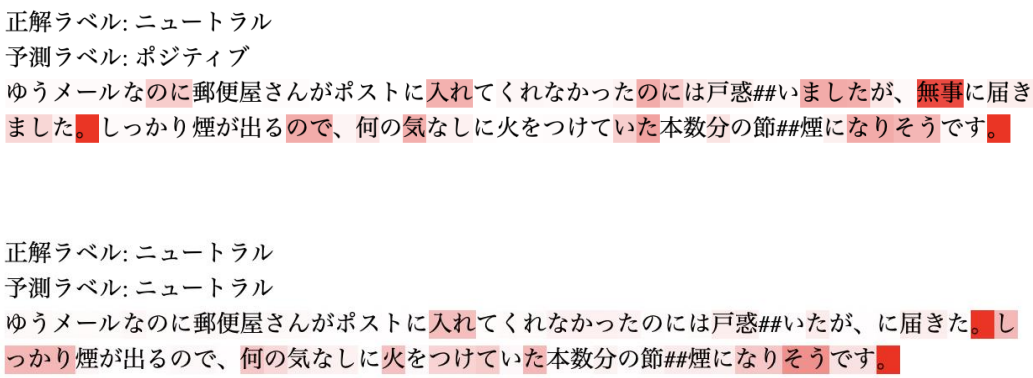

図 5.1　実験 2 でストップワードの除去を行う前後での Attention の比較

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 5.1 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 倍に増えている．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．予測を誤る要因となる単語を十分に除去できなかったのではないかと考えられる．

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．

5.2.3　実験 3．楽天データセット：使い道 6 クラス分類

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．

結果・考察

表 5.3 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 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

表 5.3　実験 3 の分類精度と t 検定を行った際の p 値

$p(w)$	手法	精度	p 値
なし	-	0.5412	-
不正解データの Attention	ストップワードリスト	0.5000	9.143×10^{-5}
	確率的ランダム除去	0.4883	2.901×10^{-7}
Attention の差	ストップワードリスト	0.5413	9.905×10^{-1}
	確率的ランダム除去	0.5475	4.170×10^{-1}
正規化した Attention の差	ストップワードリスト	0.5395	7.686×10^{-1}
	確率的ランダム除去	0.5476	2.946×10^{-1}

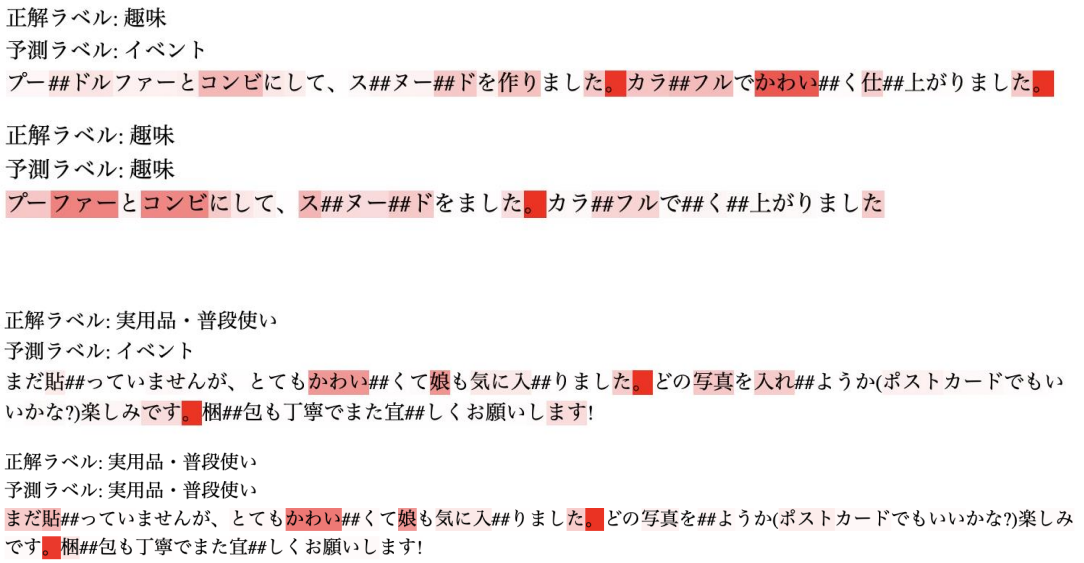

図 5.2　実験 3 でストップワードの除去を行う前後での Attention の比較

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 5.2 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 5.2 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った方が良いといえる．

5.2.4　実験 4．楽天データセット：目的 8 クラス分類

データ内容

データセットの目的の項目をラベルとして使用した．目的の項目には,「自分用」,「家族へ」,「仕事関係へ」などのラベルが付与されている．付与されたラベルを用いて 8 クラスの分類を行った．

結果・考察

表 5.4 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，ストップワードの除去を行わない場合に最も精度が高いことが分かる．ストップワードの除去を行ったものの中では，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った場合の精度

表 5.4　実験 4 の分類精度と t 検定を行った際の p 値

$p(w)$	手法	精度	p 値
なし	-	0.4992	-
不正解データの Attention	ストップワードリスト	0.4413	2.568×10^{-9}
	確率的ランダム除去	0.4344	1.539×10^{-14}
Attention の差	ストップワードリスト	0.4741	2.734×10^{-6}
	確率的ランダム除去	0.4886	8.138×10^{-4}
正規化した Attention の差	ストップワードリスト	0.4966	2.972×10^{-1}
	確率的ランダム除去	0.4941	8.036×10^{-2}

が最も高いことが分かる．

t 検定を行った際の p 値を確認すると，不正解データの Attention に着目してストップワードの生成を行った場合と Attention の差に着目してストップワードの生成を行った場合に p 値が 0.05 を下回っていることが分かる．どちらの場合にもストップワードを使用せずに分類を行った場合と比べて精度が低いことが分かる．楽天データセットを用いた目的の分類においては，不正解データの Attention に着目したストップワードと Attention の差に着目したストップワードは分類精度に悪影響を与えてしまうといえる．

ストップワードの除去を行ったものの中で最も精度が高かったのは，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った場合であった．Attention の差の値について正規化を行うことによって，ほとんどの単語で $p(w)$ が増加しストップワードとして除去されやすくなった．正規化した Attention の差の値に着目した場合には有意な差が認められなかったため，精度の低下がわずかであったといえる．Attention の差の値に着目した場合には予測を誤る要因となる単語が十分に除去できていなかったと考えられる．正規化した Attention の差の値に着目することによって，Attention の差の値に着目し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たため精度の低下が小さくなったと考えられる．しかし，正規化した Attention の差の値に着目してストップワードを用いた場合でも，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低いことが分かる．正規化した Attention の差の値に着目しても予測を誤る要因となる単語を十分に除去できていないと考えられる．

楽天データセットを用いた目的の分類において，今回提案したストップワードの生成方法は分類精度の向上に有効とは言えなかった．正規化した Attention の差に着目してストップワードリストを使用してストップワードの除去を行った場合に精度の低下が最も小さかったことが分かった．

5.3　実験 5．Twitter 日本語評判分析データセット：3 クラス分類

データ内容

Twitter 日本語評判分析データセット [13][‡]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．

結果・考察

表 5.5 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

[‡]https://www.db.info.gifu-u.ac.jp/sentiment_analysis/

表 5.5　実験 5 の分類精度と t 検定を行った際の p 値

$p(w)$	手法	精度	p 値
なし	-	0.7297	-
不正解データの Attention	ストップワードリスト	0.6738	7.929×10^{-6}
	確率的ランダム除去	0.6983	3.372×10^{-3}
Attention の差	ストップワードリスト	0.7364	4.587×10^{-1}
	確率的ランダム除去	0.7307	9.126×10^{-1}
正規化した Attention の差	ストップワードリスト	0.7360	5.356×10^{-1}
	確率的ランダム除去	0.7326	7.646×10^{-1}

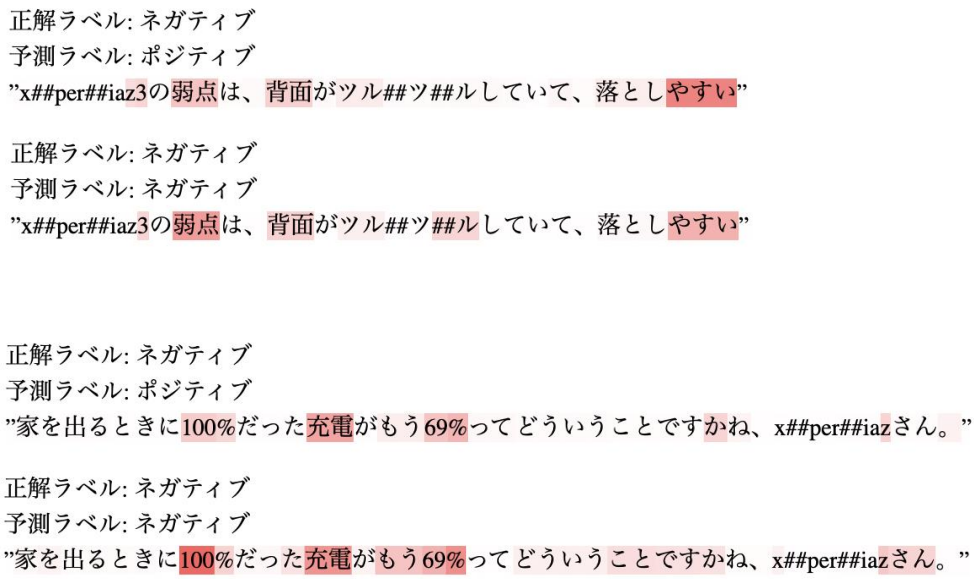

図 5.3　実験 5 でストップワードの除去を行う前後での Attention の比較

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．Attention の差の最大値は 0.496 であったため，最大値が 1 となるように正規化を行うことによって，それぞれの単語の $p(w)$ の値はおよそ倍になると思われる．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 5.3 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 5.3 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を行う場合には，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う方が良いといえる．

5.4 実験 6．livedoor ニュースコーパス：9 クラス分類

データ内容

livedoor ニュースコーパス[§]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の収集を行った．

[§] http://www.rondhuit.com/download.html#ldcc

結果・考察

表 5.6 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．livedoor ニュースコーパスを用いた分類でも不正解データの Attention 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 の差の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．

表 5.6　実験 6 の分類精度と t 検定を行った際の p 値

$p(w)$	手法	精度	p 値
なし	-	0.8971	-
不正解データの Attention	ストップワードリスト	0.8971	9.999×10^{-1}
	確率的ランダム除去	0.8911	7.238×10^{-2}
Attention の差	ストップワードリスト	0.8969	9.468×10^{-1}
	確率的ランダム除去	0.9013	2.343×10^{-1}
正規化した Attention の差	ストップワードリスト	0.9016	2.583×10^{-1}
	確率的ランダム除去	0.8984	7.471×10^{-1}

livedoor ニュースコーパスを用いたニュースカテゴリの予測を行う場合，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のが良いといえる．

第 6 章　おわりに

本研究では，BERT を用いた文書分類タスクの精度を向上させることを目的として，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 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．

正解データに出現した場合の Attention の値が小さい単語は，分類器が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．そのため，不正解データに出現した場合の Attention の値が大きく，正解データに出現した場合の Attention 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．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．

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が有意なものか確認するために統計的検定を行った．

実験の結果，今回行った六つの実験のうち四つの実験において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．

今回行った六つの実験のうち五つの実験では，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精度が低下したことが分かった．さらに．精度の変化に有意な差があることを確認した．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 2，実験 3，実験 6 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，実験 4 においては正規化を行う前に比べて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．Attention に着目した他の方法として，各単語の Attention の平均値との差に着目する方法や，平均値ではなく各単語の Attention の値を用いる方法が挙げられる．また，Attention に着目すると同時に，単語の出現頻度や文字数といった他の特徴を考慮する方法が考えられる．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

本研究を進めるにあたって，多くの方々に支えていただきました．

指導教員である鈴木優准教授には，研究の着想を得ることや，実験及び論文制作を進めることにおいて，たくさんのご指導ご鞭撻を賜りました．本研究では，なかなか良い結果を得ることができず，研究が行き詰まることもありました．そんな時には，次の実験でどのようなことをすると良さそうかご助言をいただきました．また，突然面談をお願いすることも多々あったと思いますが，いつも快く応じてくださいました．ありがとうございました．

事務補佐員の佐野さん，井尾さんには，様々な手続きを行うにあたって，お世話になりました．そのため，研究活動を滞りなく進めることができたと思います．ありがとうございました．

鈴木研究室に所属している皆様には，本研究を進めるにあたって，参考になる多くのご意見をいただきました．普段から，研究のことについて相談することで，研究がより良いものになったのではないかと思います．また，日頃の雑談などを通じて配属当初に比べて，かなり仲を深めることができたと思います．鈴木研究室に所属している皆様のおかげで，良い環境で研究ができたと思います．ありがとうございました．

国立情報学研究所からは楽天データセット, 株式会社ロンウィットからは livedoor ニュースコーパス，鈴木研究室からは Twitter 日本語評判分析データセットを提供していただきました．データセットを提供していただいたおかげで，研究を進めることができました．ありがとうございました．

本論文を無事に書き終えることができたのは，支えてくださった皆様のおかげです．心より感謝申し上げます．

最後に，大学を卒業するまでの間，経済的にも，精神的にも私のことを支えてくれた家族に対しても，感謝の意を表したいと思います．

参考文献

[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] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] 大島裕明, 中村聡史, 田中克己. Slothlib: web サーチ研究のためのプログラミングライブラリ. *DBSJ letters*, Vol. 6, No. 1, pp. 113–116, 2007.

[7] 松田寛. Ginza-universal dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

発表リスト

[1] 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』，第 175 回データベースシステム・第 148 回情報基礎とアクセス技術合同研究発表会，2022
[2] 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』，東海関西データベースワークショップ，2022
[3] 桑原悠希，鈴木優『Attention に着目したストップワードの自動生成』，第 15 回データ工学と情報マネジメントに関するフォーラム，2023