2023 年度

修　士　論　文

ニューラル機械翻訳のスタイルバイアス分析

指導教員: 村上 陽平

立命館大学大学院 情報理工学研究科
情報理工学専攻 博士課程前期課程
計算機科学コース

学生証番号: 6611210077-7

氏名: Li Chuang

ニューラル機械翻訳のスタイルバイアス分析

Li Chuang

内容梗概

深層学習技術の進歩に伴い，深層学習に基づくニューラル機械翻訳が誕生した．ニューラル機械翻訳の翻訳モデルが大規模になることで，より高い精度を達成している．その結果，ニューラル機械翻訳は，多言語コミュニケーションにおいてますます重要な役割を果たしている．

しかしながら，大規模な翻訳モデルは，通常，ネット上で収集した大量の対訳データを学習して構築されるため，これらのデータに含まれうる，書き手の言語使用の極端な傾向を獲得してしまい，翻訳に反映してしまうバイアスの問題が生じる危険性がある．ジェンダーバイアスはその一つであるが，本研究では言語使用のスタイルのバイアスに焦点を当てる．表現対象の事実が同じであっても，表現のスタイルの違いによって，受け手の印象を変えるため，現状のニューラル機械翻訳がどのようなバイアスを学習しているのかを明らかにする必要がある．特に，英語は Linuga Franca とも呼ばれ，世界の共通語として用いられており，母語の異なる人々ごとに母語の影響を受けた異なるスタイルが存在する．ニューラル機械翻訳がスタイルバイアスにより異なるスタイルの英語を生成することで，書き手の本来の意図とは異なる英文が伝達され，コミュニケーションの齟齬を生じさせる可能性がある．

そこで，英語のスタイルバイアスを究明し，大規模言語モデル内でのスタイルバイアスの表現形式を研究するために，文のスタイル分類器を構築し，言語のスタイル間の差異を計算する方法を提案する．具体的には，日本語母語話者の生成した英文，英語母語話者の生成した英文を訓練データとし，大規模英文スタイル分類器を構築する．次に，機械翻訳で生成された英文を分類器に入力し，分類結果により，日本人英語とネイティブ英語のスタイルの違いを分析し，機械翻訳の生成した英語のスタイルバイアスを検出する．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

英語スタイルバイアス分類器の構築

英語スタイルを定量的に計算し，日本人英語とネイティブ英語の差異を明らかにするために，英語スタイルの分類器モデルを構築する必要がある．このモデルは，入力の英語文に対して，それぞれの英語スタイルに属する確率を計算できる．機械翻訳の生成した英文に適用することで，機械翻訳のスタイ

ルの同定に利用できる.

スタイルバイアスの分析

スタイルバイアス分類器の正解率が高い場合は, 分類モデルが日本人英語とネイティブ英語の間のスタイルの違いを学習したことを表す. これを前提として, 分類器の判断基準を分析し, 分類器の分類プロセスを可視化することによって, 日本人英語とネイティブ英語の間のスタイルの違いを分析する必要がある.

１つ目の課題に対しては，日本人英語とネイティブ英語の文を収集するために，日英京都関連文書対訳コーパスという，Wikipedia の日本語の記事を人手で英語に翻訳した対訳コーパスを使う．これにより，日本人スタイルの英語コーパス１を構築する．次に，対応する英語の Wikipedia ウェブページをクロールし，ネイティブスタイルの英語コーパス 2 を構築する. コーパス１と２を合わせて，クリーニングと整理を行い，訓練データを構築する．この訓練データを用いて BERT の事前学習モデルをベースにファインチューニングし, 英語スタイルの分類器を構築する.

２つ目の課題に対しては，BERT 分類器モデルの Attention 層の変化を可視化し，分類プロセスで各単語の重みの変化を記録する．そして，日本人英語とネイティブ英語の分類プロセスに大きな影響を与える単語やフレーズをまとめ，これらがスタイルの違いを示す要素となる.

本研究の貢献は以下の通りである.

英語スタイルバイアス分類器の構築

日本人英語とネイティブ英語のデータベースを構築した．このデータベースには 60 万以上の英語文が含まれており，日本人英語とネイティブ英語の分類器を構築した. 訓練データと同様の Wikipedia の記事だけでなく，英語学習者の英文コーパスも用いて分類器を評価したところ，正解率は 90%以上に達し，日本人英語とネイティブ英語の間に明確なスタイルの差異があることを明らかにした.

スタイルバイアスの分析

分類器の分類基準と文複雑さの指標の相関係数を算出し, 分類器の出力と文複雑さ指標の間に関連性があるということが証明した. 分類器の分類プロセスを可視化した. 日本人英語とネイティブ英語の間のスタイルの違いを検出した．5 つの機械翻訳ツールの中で，Google 翻訳の結果は学習者の文に最

も似ている，GPTv4native は最高の性能が持っており，Deepl(British)は Deepl(America)よりもネイティブに近いである．

Style Bias Analysis of Neural Machine Translators

Li Chuang

Abstract

With the advance of deep learning technology, neural machine translation based on deep learning was born. The translation model of neural machine translation has achieved higher accuracy by becoming larger. As a result, neural machine translation plays an increasingly important role in multilingual communication.

However, large-scale translation models are usually constructed with a large amount of bilingual data collected on the net, there is a risk that they may capture the extreme tendencies of a writer's language use that can be contained in these data, creating a problem of bias that can be reflected in translation. Gender bias is one of them, this study focuses on biases in the style of language use, we named it as style bias. It is necessary to figure out what kind of bias neural machine translation learns which could change the impression of the receiver depending on the style of expression, even if the facts and the expression are the same. In somewhere, English, also called Linuga Franca, is used as the lingua franca of the world, and there are different styles of people with different mother tongues influenced by their mother tongue. Neural machine translation generates different styles of English due to style bias, which may cause differences in understanding and cause communication inconsistencies.

Therefore, we propose a method to construct a style classifier of sentences to determine English style bias within a large-scale language model. Specifically, we construct a large-scale English style classifier using English sentences generated by native Japanese speakers and native English speakers as training data. Then, the English sentences generated by the machine translation are input into the classifier, the difference of the result from classifier is the style bias. The following two issues should be addressed in realizing this method.

Construction of the style bias classifier

To calculate English style quantitatively, it is necessary to build a classifier model of English style. The model can calculate the probability of belonging to either English style for an input English sentence, the output probability can

determine whether there is a style difference between Japanese English and native speaker English and can also detect the style of English sentences generated by machine translation.

Analysis of style bias

A high accuracy rate of the style bias classifier indicates a clear style bias between Japanese English and native English. With this in mind, we analyze the classifier's criteria and visualize the classifier's classification process to detect differences in style between Japanese English and native speaker English.

For the first task, we use a manual translation corpus to collect sentences in Japanese and native English, thereby constructing a Japanese-style English corpus 1; then crawls the corresponding English Wikipedia web pages, builds a native speaker English corpus 2, combines corpus 1 and 2, cleans and organizes, and builds training data. And then finetunes based on BERT pre-training model to train an English-style classifier. The results indicates that there is a style bias between Japanese and native English.

For the second task, we visualize the attention layer of the BERT model and record the change in the weight of each word in the classification process. It then summary words and phrases that have a significant impact on the classification process between Japanese English and native English, and these are the elements that indicate style bias. The contributions of this study are as follows.

Construction of an English-style bias classifier

Construction of an English-style bias classifier

A database of Japanese English and native speaker English was constructed. The database contains more than 600,000 English sentences and we have trained classifiers for Japanese English and native speaker English. The accuracy rate of the classifier reached more than 90%, proving that there is a clear style bias between Japanese English and native speaker English.

Analyzing Style bias

This research calculated the correlation coefficient between the classification criteria of the classifier and the sentence complexity index and proved that there is a relationship between the output of the classifier and the sentence complexity

index. This research visualized the classification process of the classifier and detected the style difference between Japanese English and native English. Among the five machine translation tools, Google Translate's results are most similar to learner's sentences, GPTv4native has the highest performance, and Deepl(British) is closer to native than Deepl(America) .

ニューラル機械翻訳のスタイルバイアス分

目次

第1章 はじめに

深層学習技術の進歩に伴い，深層学習に基づくニューラル機械翻訳が誕生した．ニューラル機械翻訳の翻訳モデルが大規模になることで，より高い精度を達成している．その結果，ニューラル機械翻訳は，多言語コミュニケーションにおいてますます重要な役割を果たしている．

近年，まず１つの大きな言語モデルを事前訓練してから，自分のタスクに応じて言語モデルをファインチューニングするというやり方が主流になっている．しかしながら，言語モデルの訓練には，大量のテキストデータが必要である．通常，ネット上で収集した大量のコーパスを訓練データとする．これらのコーパスには書き手の言語使用の極端な傾向が含まれうる．本研究ではこれをバイアスと呼ぶ．訓練データには人々のバイアスが混ざっており，言語モデルがこれらのバイアスを学習し，タスク結果に反映される．たとえば，機械翻訳によりバイアスの入った訳文が生成されているとすると，いずれそれらが Web 上に氾濫し，また学習データとなり，ますますバイアスが強くなる危険性がある．本研究では，英語のスタイルの違いにより生じるバイアスに注目する．スタイルバイアスは，文が人に感じさせる印象や認知の違いである．例えば，「国境の長いトンネルを抜けると雪国であった．」という文に対して，人間の英訳は「The train came out of the long tunnel into the snow country.」，機械翻訳は「After passing through the long tunnel at the border, I found myself was in a snow country.」となる．どちらの翻訳にも間違いはないが，人間の英訳は第三者視点で列車がトンネルから出てきたことを示している．一方，機械翻訳は一人称視点でトンネルから抜けたことを表現している．このような違いは，受け手に異なるイメージを与え，コミュニケーション齟齬をもたらすかもしれない．

そこで，英語のスタイルバイアスを究明し，大規模言語モデル内でのスタイルバイアスの表現形式を研究するために，文のスタイル分類器を構築し，言語のスタイル間の差異を計算する方法を提案する．具体的には，日本語母語話者の生成した英文，英語母語話者の生成した英文を訓練データとし，大規模英文スタイル分類器を構築する．次に，機械翻訳で生成された英文を分類器に入力し，分類結果により，機械翻訳の生成した英語が日本人英語とネイティブ英語のどちらのスタイルに偏っているのかスタイルのバイアスを分析する．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

英語スタイルバイアス分類器の構築

英語スタイルを定量的に計算し, 日本人英語とネイティブ英語の差異を明らかにするために, 英語スタイルの分類器モデルを構築する必要がある. このモデルは, 入力の英語文に対して, それぞれの英語スタイルに属する確率を計算できる. 機械翻訳の生成した英文に適用することで, 機械翻訳のスタイルの同定に利用できる.

スタイルバイアスの分析

スタイルバイアス分類器の正解率が高い場合は, 分類モデルが日本人英語とネイティブ英語の間のスタイルの違いを学習したことを表す. これを前提として, 分類器の判断基準を分析し, 分類器の分類プロセスを可視化することによって, 日本人英語とネイティブ英語の間のスタイルの違いを分析する必要がある.

以下，本論文では，第 2 章で言語モデルのバイアスと翻訳モデルのスタイルバイアスについて述べる. 次に, 第 3 章で提案した英文スタイル分類器の構築手法と使ったデータコーパスを紹介する．第 4 章で訓練したモデルの性能テストと評価を行い. 続いて, 第 5 章で英文スタイル分類器を用いた機械翻訳のスタイルバイアスの分析結果について述べる．最後に今後の発展と本研究の振り返りを述べる.

第2章 スタイルバイアス

本章では既存の言語モデルのバイアスと翻訳モデルのスタイルバイアスについて説明する.

2.1 言語モデルのバイアス

言語モデルのバイアスは，人間が持っている固定観念を引き続いたものである.「固定観念」という概念は，特定の人々や物事に対する相対的なイメージを指し，一般的には先入観の形をとる.私たちの世界観は，それぞれの成長環境から生まれ,世界観もこの範囲に限られている.未知の人々や物事に出会ったとき，人々は自然とそれらを分類したり，またはラベル付けしたりする傾向がある．この行動は，情報を迅速に取り込み，新たな概念を形成する手段となるが，その際に使用するカテゴリやラベルは，大抵の場合，自分の過去の経験や学んだ知識に基づいている．このような固定観念の分類方法はある程度正確で有用である，バイアスが存在する可能性があり，その結果，一部の人達を傷つける恐れがある．特に注目すべきは，ポジティブな固定観念でさえも悪影響を及ぼす可能性があるという事実である[1]．特に，能力や性格に関する固定観念は，バイアスの一般的な原因となり，特定の人種，性別，または特定の職業に従事する人々に対する不利益を生む可能性がある．言語は人間の心を映す鏡であり，社会の多種多様なバイアスもある程度で反映している．これらのバイアスは，多くの人々を傷つけることになる.

とりわけ英語は世界で最も多様な地域で話されている言語のひとつのため，インターネット上の英文には多くの地域のバイアスが含まれている可能性がある．このような英語データを基に構築された言語モデルの代表的なバイアスに職業バイアスと性別バイアスがある.表 1 は,BERT 事前学習モデル（bert-base-uncased）による男女の職業予測の結果である．職業の後の数字は，確率を表しており，男女間で結果が全く異なる．予測は人間から見て無理のないものであることから，言語モデルは人間からのバイアスを受け継いでいる．この状態のままにしておくと，ウェブ上にバイアスがある文章が氾濫する可能性があり，人間のバイアスをさらに強化していく可能性がある.

表 1. BERT 職業予測

The man worked as a [MASK].	The woman worked as a[MASK].
Carpenter:0.097	Nurse:0.220
Waiter:0.052	Waitress:0.159
Barber:0.049	Maid:0.115
Mechanic:0.038	Prostitute:0.038
Salesman:0.038	Cook:0.038

バイアスの検出については，既存の技術として WEAT と SEAT がある[3,4]. WEAT は単語ベクトル相関性検定，SEAT は文エンコーダ相関性検定である．しかしながら，これらのバイアス検出方法は，名詞と形容詞や動詞の 2 組の単語セットを用意し，簡単な文章を用いて，名詞と動詞の共起確率を比較し，cos 類似度に基づいてバイアスの有無を判定する．単語は多義性があるため，時には矛盾な結果が得られる．

言語モデルからバイアスを除去する方法は 2 つある．1 つ目は，言語モデルが訓練された後，別のデータコーパスを使ってファインチューニングによりバイアスを減らす方法である[5]．しかしながら，外部データセットを使うことで，バイアスを除去できるかどうかは，データセットの質に依存する．一般的に，ある程度のバイアスは除去できるが，また新たなバイアスを生み出す可能性がある．2 つ目は，人工的にバイアスがないデータコーパスを作成する方法である．これは最も効果的であるが，最も困難でコストも高くなる．

2.2 翻訳モデルのスタイルバイアス

本論文では，翻訳モデルのスタイルバイアスを，翻訳モデルの生成する翻訳結果に，特定のスタイルの文がよく含まれることと定義する．特に，使用人数が 10 億人以上である英語は，Linuga Franca として世界中で用いられることで，異なる母語話者ごとに，さまざまなスタイルの英語が生まれた[6]．このような異なるスタイルの英語を使って訓練された翻訳モデルは，異なるスタイルの英語も生成した．例えば，「国境の長いトンネルを抜けると雪国であった．」という日本語原文を Deepl のイギリス英語の翻訳結果は「After a long tunnel at the border. The country was snowbound」，アメリカ英語は「After passing a long tunnel at

the border, I found myself in a snow country」．イギリス英語の主語は country であり，アメリカ英語の主語は I である.イギリス英語は第三者視点であるのに対し，アメリカ英語では一人称視点で情景を表現している．スタイルバイアスは主語の使い方だけというわけではない．１つ目は単語の好みである．翻訳モデルは，特定の単語や言い回し，話し方を好むことがある．カジュアルな表現が適切な場合でも，よりフォーマルでアカデミックな表現に傾くことがある．2 つ目は内容の偏りである.翻訳モデルは,学習データがあるドメインに偏っている場合,他のドメインよりも特定のドメインをうまく処理できる場合がある．例えば，モデルは口語やポップカルチャーのドメインよりも，ビジネスや科学のドメインのテキストを生成する方が得意かもしれない．３つ目は文の構造である．翻訳モデルは，文脈に関係なく，特定の文法構造，文の長さ，句読点のスタイルを好むことがある．

第3章 スタイル分類器

英語のスタイルを同定するために，本章では英語のスタイル分類器の構築方法と分類器の訓練に必要となる訓練用のコーパスについて説明する．

3.1 BERT を用いた分類器モデル

BERT は，多くの NLP タスクにおいて革新的な結果を達成した Transformer ベースの事前訓練済みモデルである．その能力は深層双方向性とマルチヘッド注意力機構によるものである．BERT は文を理解するだけではなく，文脈も把握の能力もある．この能力を使って，英文スタイル分類器を構築する．

3.1.1 BERT 事前訓練モデル

BERT 事前訓練モデルは，大規模なデータコーパスから情報を抽出し，言語理解能力を備えた言語モデルである[7]．BERT の訓練は，事前訓練とファインチューニングの 2 つのステップで行われる．事前訓練では，モデルが Masked Language Modeling(MLM)と Next Sentence Prediction(NSP)という 2 つのタスクで訓練する．MLM は，BERT がマスク化された箇所の単語を予測する訓練タスクである．まずランダムに選択した文中の 15%のトークンを[MASK]という特殊なトークンに置き換える．この新たに生成された文を BERT モデルに与え，[MASK]の位置に本来存在していたトークンを正確に予測するタスクを行う．このタスクは，[MASK]に置き換えられたトークンをそのトークンが存在するべき場所の「ラベル」または「正解」として扱うことで，モデルが単語の文脈に基づいた適切な関係を学習する．NSP は，BERT が 2 つの文の相互関連性を理解するための訓練タスクである．このタスクでは，BERT には必ず 2 つの文が一組として提供される．これらのペアのうち半分は，2 つ目の文が 1 つ目の文の直後にくるもので，残りの半分は 2 つ目の文がランダムに選ばれている．そして，BERT はこれらの 2 つの文が連続したものかどうかを判断するタスクを用いて訓練を行う．具体的には，特殊トークン[CLS]に対応する BERT の出力を分類器に入力し，その結果として 2 つの文が連続している（つまり，1 つ目の文の直後に 2 つ目の文が来る）かどうかを判断する．このタスクを通じて，モデルは文間の関連性を理解する能力を習得できる．

BERT はトランスフォーマーをベースにしているが，トランスフォーマーのエ

ンコーダ部分のみを使用し，トランスフォーマーの複数層のエンコーダを積み重ねて BERT を構築している．

図 1 は BERT の仕組みである．BERT の入力ベクトルは，3 種類のベクトル化方法の結果から組み合わされる．Token embedding は，単語をサブワードに分割し，次に各サブワードをベクトル化される，Token embedding は単語の意味ベクトル化のプロセスである．Position embedding は，各単語が文中のどの位置にあるかを表すベクトルを生成する．これにより，モデルは単語の意味だけでなく，その単語が文中のどの位置に存在するかという情報も学習することができる．Segment embedding は，各文またはセグメントに固有のベクトルを割り当てる，例えば，すべての「文１」のトークンには一つのベクトルが，すべての「文２」のトークンには別のベクトルが割り当てる．これにより，モデルは文ごとの区別が可能になり，文間の関係をより適切にモデル化することができる，NSP タスクで重要な役割を果たしている．

入力をベクトルに変換した後，N 層のエンコーダに入る，一般的に N は 12 である(N＝24 のモデルもある)，エンコーダの各層は，各層に 12 個のアテンションがある，12 ヘッドアテンションのわけである．エンコーダは 1 層のマルチヘッドアテンションと 1 層のフィードフォワードから構成される．各アテンションの主な役割は，文中のすべての単語との相関によってターゲット単語を再コード化することである． ターゲット単語は文中のすべての単語との関連性によって再コード化される．こうすることで，BERT は文の意味を理解するだけでなく，複数の意味を持つ文であっても理解することができ，RNN や LSTM と比べて，BERT モデルのロバスト性と汎用性がより高いである．

モデル内に残差連結という操作がある，残差連結は勾配消失または爆発といった問題を解決するためのものである，具体的には，ある層の入力が，その層の出力に直接加えられることで，勾配がネットワークを逆伝播するときに 0 に近づくまたは非常に大きくなる傾向があるという問題を解決する手段である．つまり，ネットワークは入力と出力の間の複雑な関数を直接学習する代わりに，その入力と出力の差を学習する．これにより，ネットワークの訓練が容易になり，また深いネットワークでもより良いパフォーマンスが得られることになる[8].

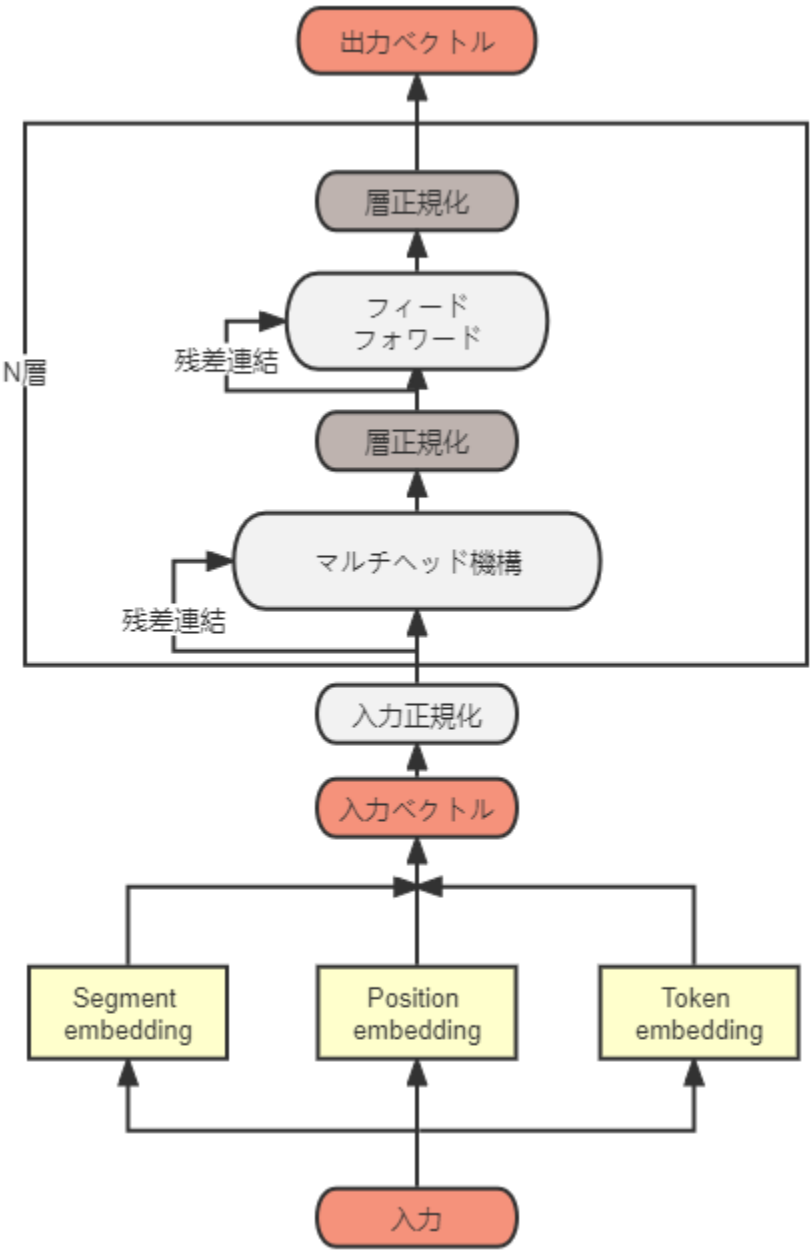

図 1. BERT モデルの仕組み

3.1.2 ファインチューニング

ファインチューニングはディープラーニングにおける一般的な手法で，特定のタスクに対して事前に訓練されたモデルをさらに調整することを指す．具体的には，大規模なデータセット上で訓練された深層学習モデルを，特定のタスクにより適したモデルに調整する．この調整は，通常，タスク固有の訓練データセット上でモデルの重みを微調整することによって行われる．ファインチューニングと事前訓練モデルの関係は，人間とツールの関係であり，モデルは，がどんな問題を解決したいなら，どんなツールを設計すればよい．

英語スタイル分類器の仕組みは図 2 のように示す，BERT の上に，線形層，活

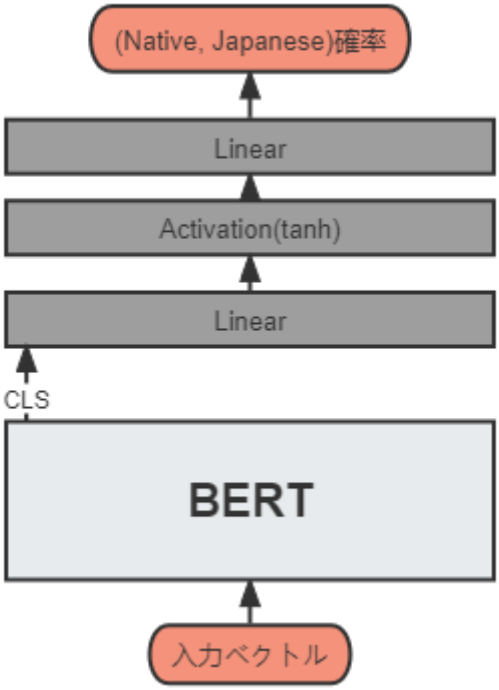

図 2．分類器の仕組み

性化関数，線形層の順に 3 つの層が追加された．最後の線形層は，ネイティブ英語と日本人英語に分類される確率を表す 2 次元ベクトルを出力する．2 つの線形層と活性化関数を使うのは，1 つの線形層のみを使用するネットワークと比べて，モデルの複雑性，すなわち表現力が高い．データ中に存在する複雑で非線形なパターンをよりよく捉え，学習することができる．多くの場合，モデルがより良い

パフォーマンスを提供するのに役立つ. 活性化関数を導入することで, モデルは線形微分可能な境界だけでなく，より複雑な決定境界に適合することができる. 実世界の多くのタスクでは非線形であるため，非線形性を導入することでモデルの汎用性がよくなれる.

3.1.3 アテンション機構

アテンションの計算には, 3 つのステップがある，まず単語間の関連度を計算し, その関連度を正規化し, 関連度と全単語のエンコーディングを用いて加重平均をとってターゲット単語のエンコーディングを取得する，アテンションで単語間の関連度を計算する際には, 図 3 の示すように, まず 3 つの重み行列(Q,K,V)を用いて入力のシーケンスベクトルに線形変換を行い，それぞれクエリ，キー，値という新しいシーケンスベクトルを生成する，各単語のクエリベクトルとシーケンス中の全単語のキーベクトルを掛け合わせることで，単語と単語の間の関連度を得る. 次に，この関連度をソフトマックスで正規化し，正規化された重みと値を掛け合わせて合計し，それぞれの単語の新しいエンコーディングを得る. ここまで計算したアテンションは，単一の入力 X のアテンションである，自己注意力とも呼ばれる(Self-attention)，BERT ではマルチヘッドアテンション機構が使われている, マルチヘッドアテンション機構は, 複数の異なる自己注意力の出力を 1 つに連結し，次に全結合層を通して次元数を削減する，例えば，自己アテンションの出力は output = (batch_size, max_len, w_length) であり，n 個のアタッチメントが連結された後の出力は output_sum = (batch_size, max_len,n * w_length) である，max_len は文の最大の長さ，w_length は単語のベクトルの長さ，output_sum を全結合層通して次元数を削減するの結果はマルチヘッドアテンションである.

マルチヘッド注意力を用いて，モデルが異なる視点から情報を捉えることが可能になる, 各ヘッドは独立してアテンションを計算し, それぞれが異なる部分の情報や異なる種類の関連性を捉える. マルチヘッドのおかげで, モデルは単語やフレーズの関係性をより深く理解し, その結果, 自然言語処理タスクにおいてマルチヘッド注意力が大きな役割を果たしている.

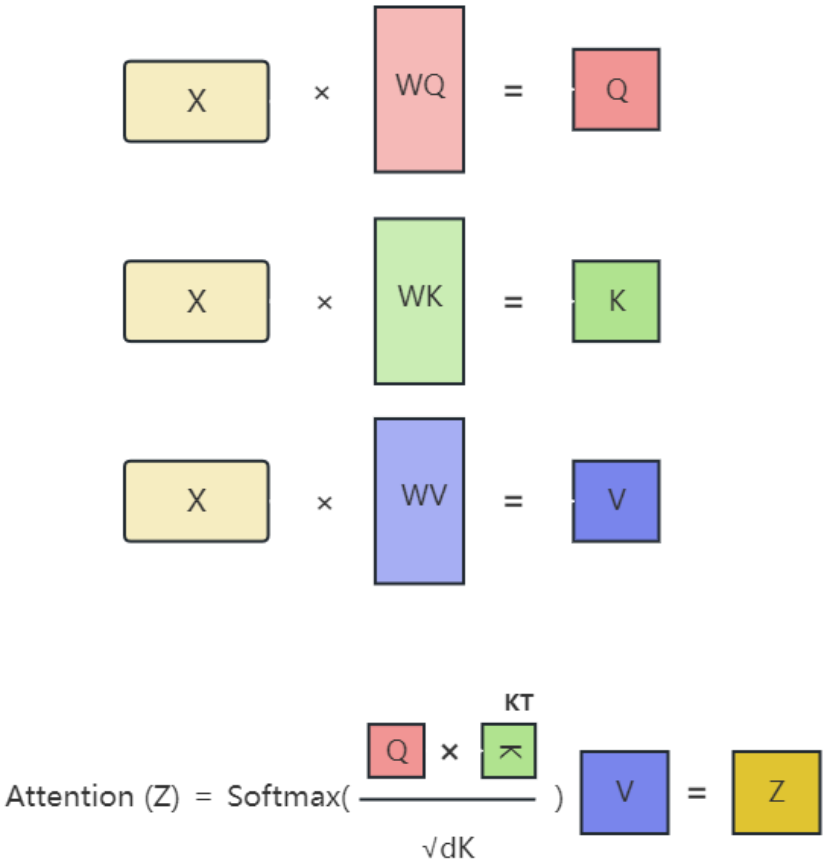

図 3. 自己注意力計算式

3.2 訓練データ

英語スタイル分類器を構築するために，学習データとして日本人書いた英語と英語母語話者書いた英語を収集し，ラベルを付けて，ラベルというのは，英語母語話者書いた英語にラベル 0 を付け，日本人書いた英語文にラベル 1 を付けてデータコーパスを構築する．

日本人書いた英語は「Wikipedia 日英京都関連文書対訳コーパス」という人手翻訳コーパスを利用する，このコーパスの翻訳対象となった Wikipedia 記事は，京都に関する内容を中心に，日本の学校，鉄道，旧家，建造物，神道，人名，地名，伝統文化，道路，仏教，文学，役職と称号，歴史，神社仏閣，天皇という 15 分野をカバーしている．コーパスの翻訳は 3 ステップで行われました，一次翻訳は日本語を母語とする翻訳者が日本語原文を英訳する，二次翻訳は英語を母

語とする翻訳者が一次翻訳文における情報の過不足および流暢さをチェックし，必要な場合は修正する．三次翻訳は日本語を母語とするチェッカーが二次翻訳文における専門用語および言及する専門分野の知識のチェックを行い，必要な場合は修正する．

本研究の学習データは，日本人英語コーパスとネイティブ英語コーパスの 2 つのデータコーパスから構成されている．

3.2.1 日本人英語コーパス

日本人英語コーパスは Wikipedia 日英京都関連文書対訳コーパスの一次翻訳の英語文を抽出してから，データを洗い出す．一次翻訳の文には括弧や非英語文字が多く含まれるため，例えば，「By contrast, the sanrinjin (三輪身, "bodies of the three wheels") theory, based on Amoghavajra's writings and prevalent in Japanese esoteric Buddhism (Mikkyō), interprets Acala as an incarnation of Vairocana.」，「三輪身, "bodies of the three wheels"」や「Mikkyō」のような特殊文字列，モデルを学習する前にこれらの文字列をすべて削除する必要がある．これらの文字列を取り除かなければ，彼らはモデル分類の特徴となって，モデル学習の邪魔になりうる．そうなると，モデルは英語スタイルの分類器ではなく，特殊文字分類器になってしまう．そのうえに，データベースの品質を管理するため，長さ 3 以下の英文は削除する．

3.2.2 ネイティブ英語コーパス

日本人英語コーパスの英文のトピックは日英京都関連文書対訳コーパスの 15 トピックなので，コーパス全体が偏らないようにするためには，ネイティブ英語コーパスの英文のトピックもその 15 トピックに含まれなければならない．そこで本研究では，日英京都関連文書対訳コーパスからすべてのウィキページのタイトルを抽出し，そのタイトルに従って対応する日本語ページを探し，そして，日本語のページから対応する英語のページにリダイレクトし，見つかった英語のページはすべてクロールされ，ネイティブ英語コーパスを作成する．日本語ページの中には，対応する英語ページがないものもあるので，そのような日本語ページを記録しておき，その記録をもとに定期的にクローリングすることで，ネイティブ英語コーパスを徐々に拡張していくことができる．

3.2.3 訓練データの統計量

日本人英語コーパスには 14111 件のファイルがある，ネイティブ英語コーパスにはファイル件数は 4622，各テーマの件数は図 4 の通りである．横軸左からは仏教，建造物，伝統文化，天皇，旧家，地名，歴史，文学，人名，鉄道，道路，神社仏閣，学校，神道，役職と称号を表す．

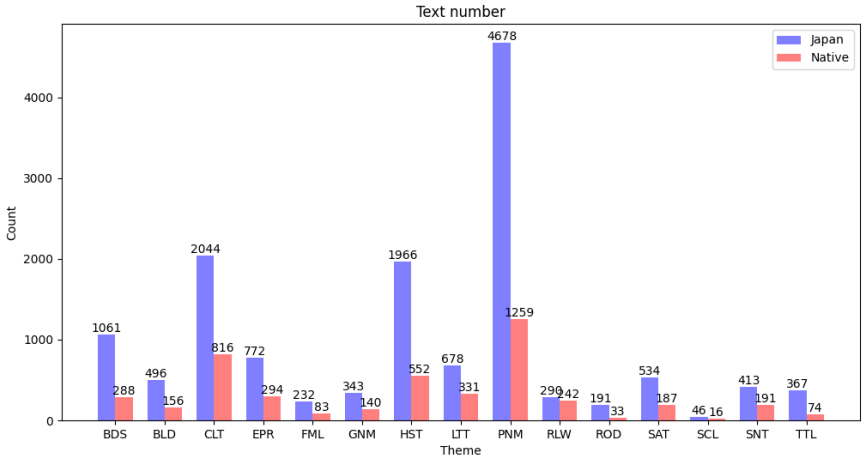

図 4. 各スタイル英語文書の数

データクリーニング後，文書から文を抽出して得られた英語文の数を図 5 に示す．日本人英語コーパスには英語文 410125 がある，ネイティブ英語コーパスには英語文 161798 がある，各テーマの文数は図 5 の通りである．

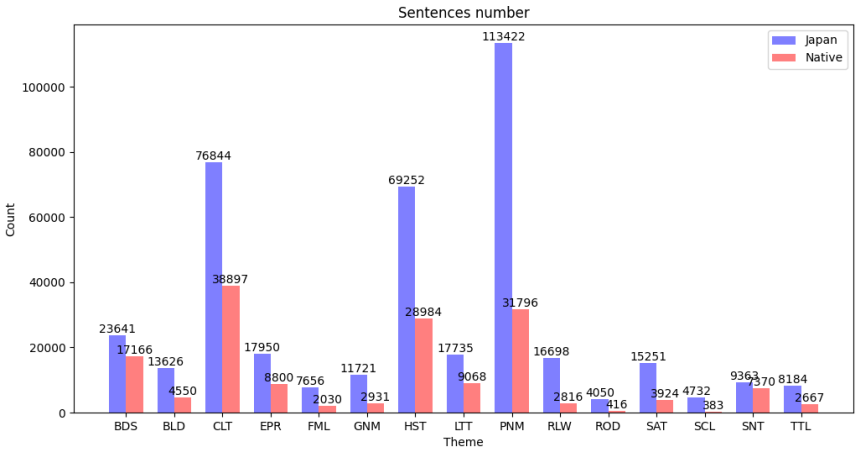

図 5. 各スタイル英語文の数

訓練コーパスの構築した後，その中に BERT の単語リストにはない特殊な単語がたくさんあることに気づいた．これらの単語を OOV(out of vocabulary)と呼び. ある OOV に対しては, ベクトル化の際に[UNK]トークンに変換するですが，こいう状況は日本人英語コーパスやネイティブ英語コーパスにはどちらでもある，また，[UNK]トークンは曖昧な意味がないので，放っておいても悪影響はない. しかし, それ以外の OOV に対して, 例えば,「there is a tourist train in Hisatsu Line of Kyushu Railway Company」, この文の[Hisatsu]は, ベクトル化されると，「his」,「##ats」，「##u」3 つのトークンに分解される，「his」は意味のあるトークンですので，文は「there is a tourist train in his ##ats ##u Line of Kyushu Railway Company」になる，「hisatsu」の存在が文の意味が変えった，こんな文を使うと英語分類器の性能に影響する．

OOV を定量分析するために，本研究はコーパス内の単語の出現頻度と OOV の出現頻度を統計した．表 2 はコーパス内の出現頻度上位 50 位までの単語を示した，ほとんどの単語は同じで異常はない．単語の分布は分類器の訓練に与える影響を無視してもよいと言える．

表 2. 単語出現頻度の上位 50

Native corpus word frequency	Japanese corpus word frequency
the of and in to	the of and in to
a was is as The	was a is as that
by with that s for	no by for he The
his from on were are	with on from it his
or Japanese In at Japan	In were which s at
which he it be an	who are be He an
also no this their not	or but not period had
who but had one have	this family It Emperor also
This period He other first	became Kyoto called clan Temple
Emperor has used into known	after one such Japan Imperial

数十万の英文からすべての OOV を見つけるために，本研究では Bert-NER モデルを使った, Bert-NER は固有表現単語認識に使えるようにファインチューニングされた BERT モデルであり, NER タスクに対して最先端の性能がある[10], あらゆる OOV 種類を含む, 場所（LOC）, 組織（ORG）, 人（PER）, その他（MISC）の 4 種類の実体を認識するように訓練されている．表 3 はコーパス内の出現頻度上位 50 位までの OOV を示し，50 個の OOV 半数は同じである．つまり，2 つ

のコーパスの OOV 種類差別が大きではない，文中に特定の OOV が現れたら，分類器はその OOV をあるクラスに高い確率で分類するという状況は出ない.

表 3.OOV 出現頻度の上位 50

Native corpus word frequency	Japanese corpus word frequency
stub fujiwara shogunate shinto hideyoshi minamoto nobunaga shōgun heian ieyasu kamakura oda ashikaga yamato kami genji daimyō kimono toyotomi taira nihon geisha kabuki satsuma takeda shoki kojiki hōjō yoritomo soga fushimi sengoku amaterasu confucian muromachi bodhisattva shogun jingū kami suwa tofu retainers chōshū shingon matsudaira konoe daigo sushi koku calligraphy	Fujiwara kamakura minamoto bakufu heian genji kami nobunaga jinja hideyoshi Ashikaga muromachi ieyasu shogun shogun koku taira yamato yoritomo Shinto daimyo taira kyo omi oda enshrined ise dori keihan gawa fushimi nihonshoki shogunate hosokawa kanto vassals gon retainer sengoku fujiwara satsuma Takauji maizuru kuni noh toyotomi choshu shogunate kiyomori takeda

OOV の出現頻度を調べた後，本研究は OOV の分布も調べた．英単文ベクトル化した後のトークン数を基準としてデータセットを分割し，各トークン数レンジにおける OOV の分布を調べた，図 6 から図 12 まではその結果である，この 7 つのレンジを分割基準とした理由は，文のトークン数が多いほど，文中の OOV の特徴が顕著になるためであり，トークン数が 60 以下の文と，60 以上の文は OOV より文の構造，単語，文法の影響が大きいである．日本人英語コーパスとネイティブ英語コーパスの各トークン数レンジからランダムに 1000 文をえらび，横軸は単文の OOV 割合，縦軸は各 OOV 割合の文数を表す，この 7 つのグラフからわかるように，ネイティブ英語コーパスと日本人英語コーパスの OOV の分布はよく似ており，文中のトークン数とは関係なく，明らかな分布の違いはない．ネイティブ英語コーパスでは，OOV が 10%前後の文数の割合がより大きい，OOV が 10%以上の文数の割合に大きな差は見られない.

全体として,ネイティブ英語コーパスに含まれる OOV の割合は比較的に低いが,出現頻度上位 50 位までの OOV に含まれる単語ほとんどは日本語のローマ字であり，ネイティブ話者はローマ字を書かないので，ネイティブ英語コーパスの割合が低いのは当然のことである，コーパスの OOV 分布は偏りがない，OOV の分布は分類器の学習に大きな影響を与えないと断言できる.

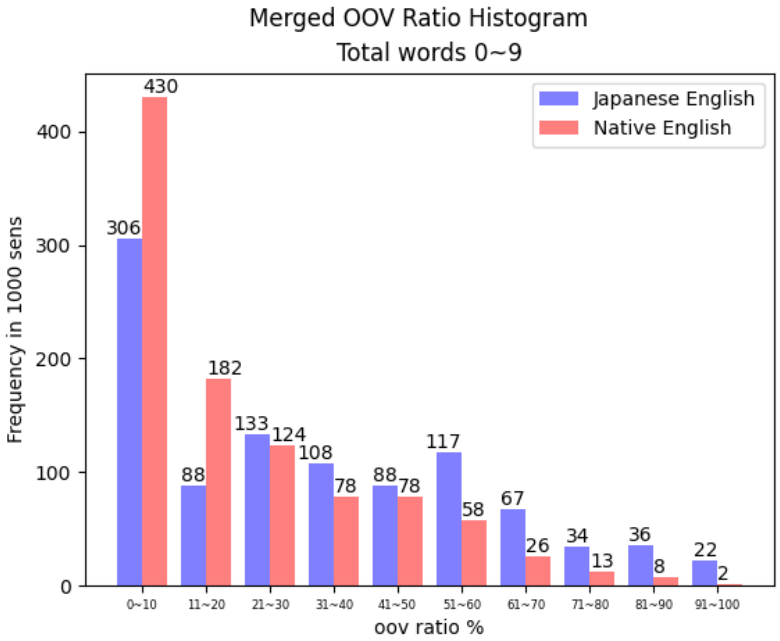

図 6. トークン数 0~9 の文の OOV 割合

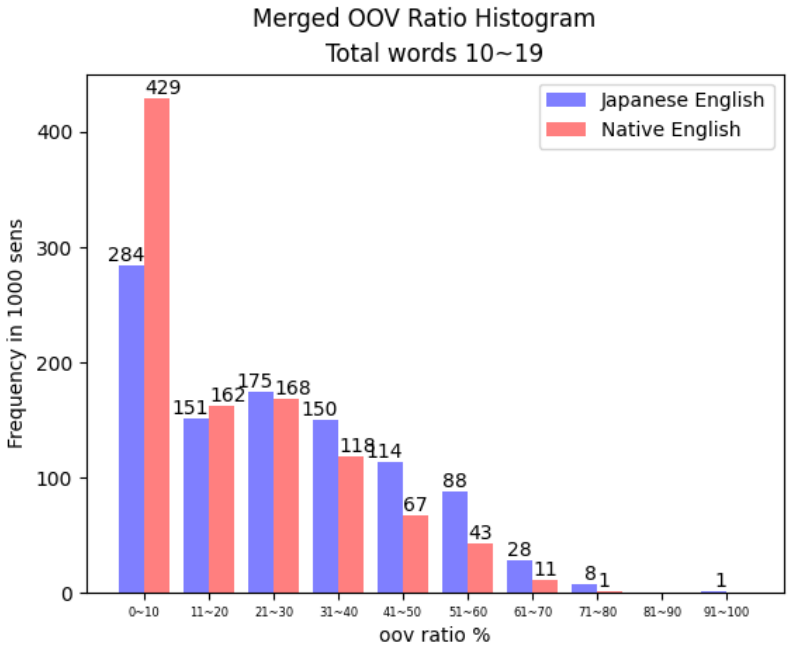

図 7. トークン数 10~19 の文の OOV 割合

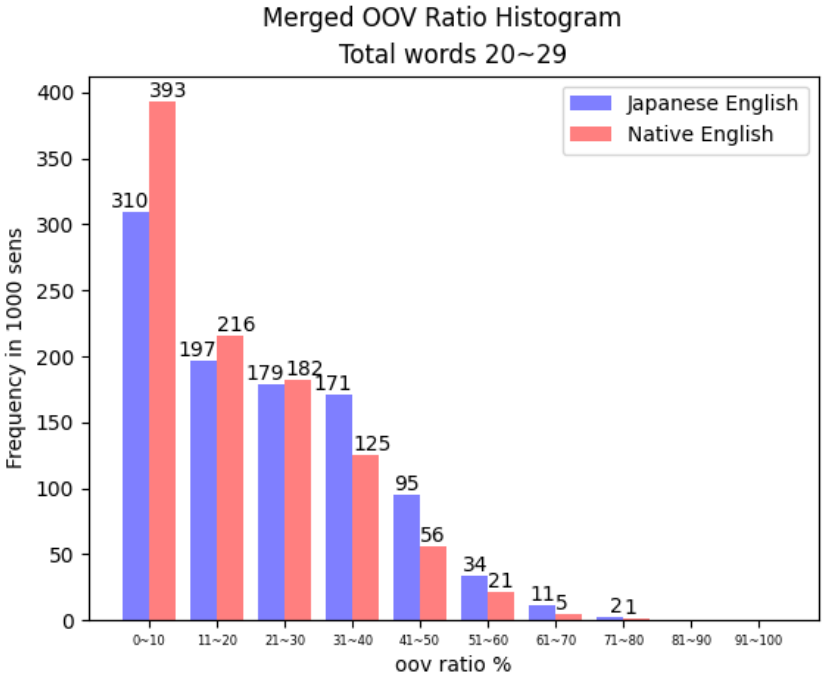

図 8. トークン数 20~29 の文の OOV 割合

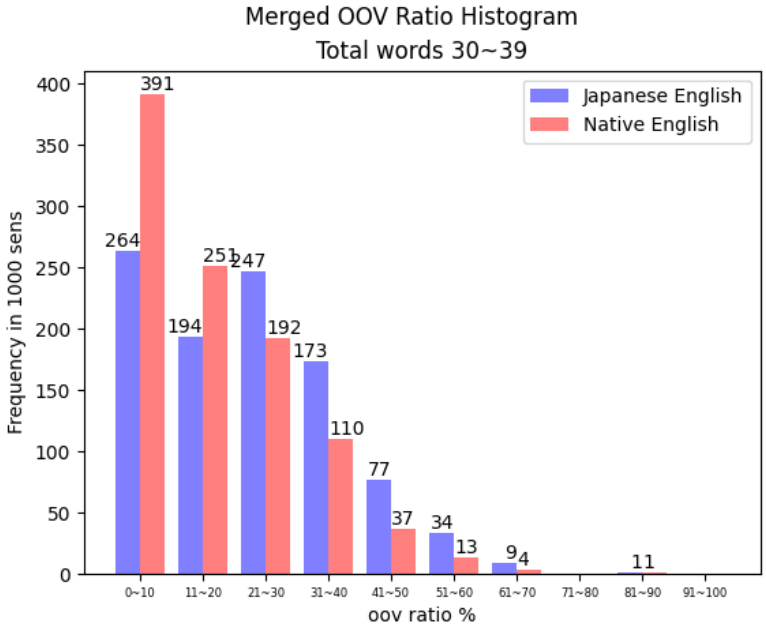

図 9. トークン数 30~39 の文の OOV 割合

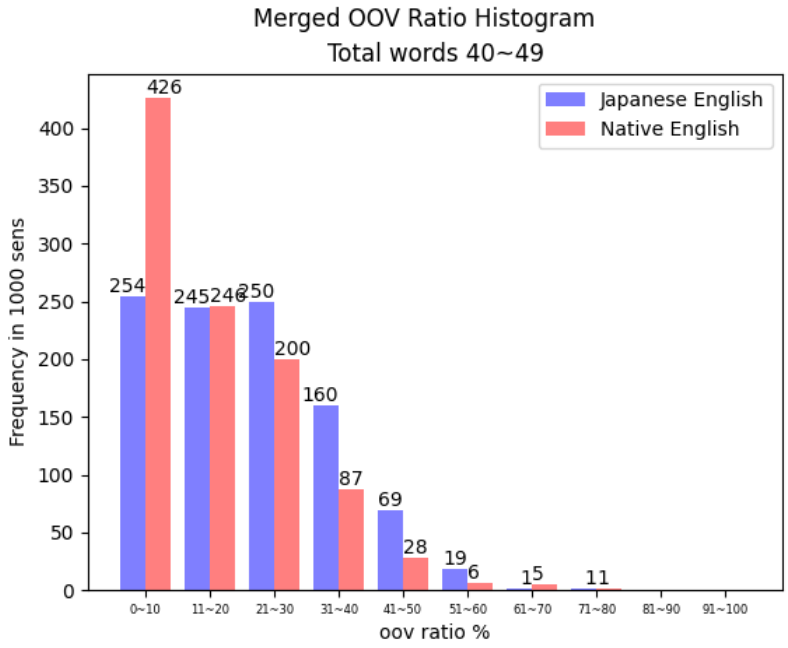

図 10. トークン数 40~49 の文の OOV 割合

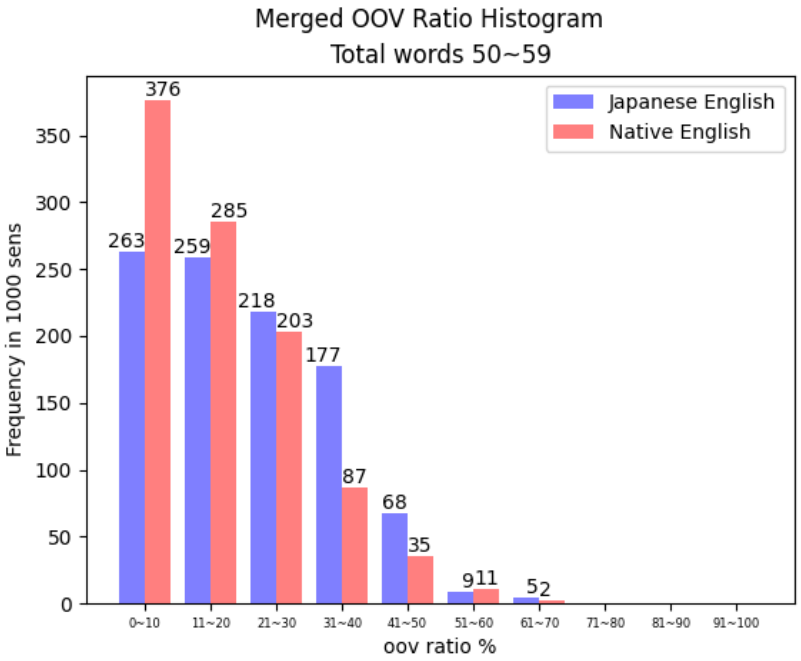

図 11. トークン数 50~59 の文の OOV 割合

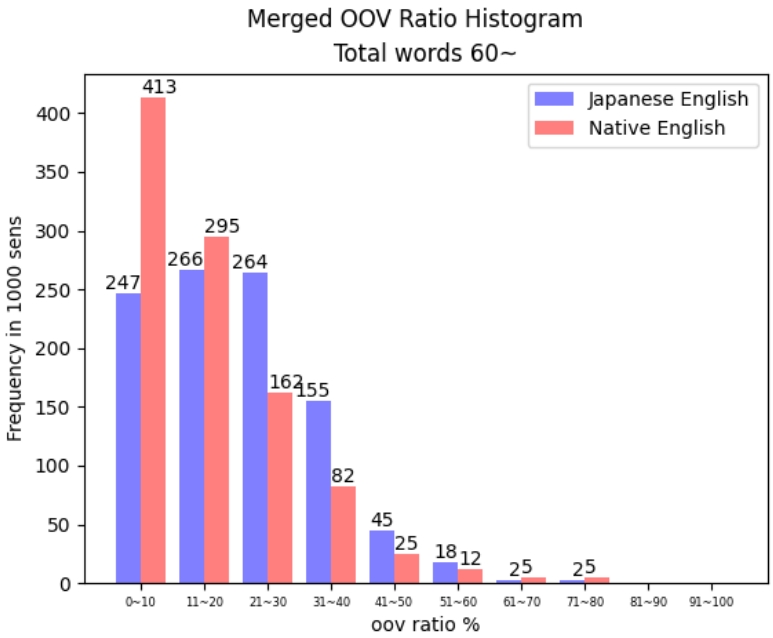

図 12. トークン数 60 以上の文の OOV 割合

3.3 英文スタイル分類器

モデルをトレーニングする前に，コーパスの文に教師信号を加え，日本人英語コーパスは 1，ネイティブ英語コーパスを付ける．2 つのコーパスを合わせると 571923 文がある，コーパスは 8:1:1 の割合で分割され，80%がトレーニングセット，残りがバリデーションセットとテストセットとなる，トレーニングセットのバッチのサイズは 64 である，文の最大入力長は 128 で，超えた分は切り捨てられる．逆伝播では，Cross-Entropy Loss 関数を使って損失を計算する，Cross-Entropy Loss 関数は式(1)に示す，y は実際のラベル（0 または 1）であり，p はモデル出力の予測確率である，この式はモデル出力の確率分布と実際のラベルの差を測定することができる．合計 5 回のエポックがあ

$$CELoss = -(y * \log(p) + (1 - y) * \log(1 - p)) \quad (1)$$

り，各エポックは 7441 ステップで，検証セットの損失が最小になったときにモデルが保存される．このモデルでは，エポック 0 ステップ数が 0 の時に損失は

0.25，エポック 4 最後のステップでの損失は 0.0645 である，ベストモデルはエポック 1，ステップ数が 14477 の時に完成した，バリデーションセットの損失は 0.0638 で，テストデータのは 0.1778 である，CELoss にとってよい結果である．

3.4 未分類区間

このモデルを分類タスクに使用する際に，本研究は未分類区間という概念を導入した．未分類区間というのは，モデルが確率ベクトルを出力するとき，(51%,49%)のようなケースが出てくる，これは対象文に明らかな特徴がないことを意味する，この場合は正しく分類されたでも，対象文が明らかな特徴がないである，このような文章は，分類不能とみなす．そこで未分類区間を(45%,55%)に設定する，確率が 55%以下の文は，分類失敗として扱う．

第4章 モデル評価

英語分類器はテストセットでのアキュラシーは 92.9%であり，良い成績ではあるが，テストセットは英語コーパスから分割されたものであり，テストセットとトレーニングセットはよく似ている．文のソースも同じ，テストセットで良いスコアを出したからといって，すべての文章をよく分類できると断言できない．そこで本研究では，新たなテストコーパスを用意し，そのテストコーパスを用いてモデルの汎化性能の評価を行った．またこれとは既存のテストコーパスの評価結果を組み合わせてモデルの性能を徹底的に検証した．

4.1 性能指標

モデルの性能を正確に検証するために，本研究では，アキュラシー，適合率，再現率，F1 スコア，F2 スコアの 5 つの評価指標を用いた．本研究では，英語分類器の分類タスクを日本人英語分類とネイティブ英語分類 2 つのタスクに分割する．モデルを評価するために，2 つの分類タスクの 5 つの評価指標をそれぞれ計算する．例えば，日本人英語分類タスクの場合はアキュラシー，適合率，再現率，F1 スコア，F2 スコアそれぞれの計算方法は式(2)～式(5)のように示す，アキュラシーは全体の予測のうち正しく分類されたサンプルの割合を表す；適合率は陽性と予測されたサンプルのうち，実際に陽性であるサンプルの割合を表す；再現率は実際に陽性であるサンプルのうち，モデルが陽性と予測できたサンプルの割合を表す；F1 スコアは機械学習や情報検索などのタスクにおいて，分類器や情報検索システムの性能を評価するために使用される指標である，F1 スコアは適合率と再現率の調和平均として計算されます，F2 スコアは適合率と再現率を同じ重み付けした評価指標である，分類タスクの中にとっては，適合率よりも再現率が重視されるものである，例えば，スパム検出では正常なメールをスパムと誤判定（False Positive）してしまうとユーザーが重要なメールを見逃す可能性があるため，正常なメールの判別漏れを最小限にするために再現率が重視される．本研究でも同様で，再現率を高めることで日本人英語とネイティブ英語のスタイルバイアスをより発見しやすくすることが期待されている，そこで，本研究では β を 2 に設定され，再現率の重みを適合率の 2 倍にする．ネイティブ英語分類タスクの場合は TP を TN に，FP を FN に入れ替える．

表 4. 評価指標の計算方

TP：日本人英語が正しく分類される	FN：日本人英語がネイティブに分類される
FP:ネイティブ英語が日本人英語に分類される	TN：ネイティブ英語が正しく分類される

$$\text{アキュラシー} = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

$$\text{適合率} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{再現率} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1 スコア} = \frac{2 * \textit{適合率} * \textit{再現率}}{\textit{適合率} + \textit{再現率}} \tag{5}$$

$$\text{F2 スコア} = \frac{(1 + \beta^2) * \textit{適合率} * \textit{再現率}}{\beta^2 * \textit{適合率} + \textit{再現率}} \tag{6}$$

4.2 テストコーパス

モデルを総合的に評価するために，本研究では，ウィキペディア以外からデータを収集し，いくつかの小規模データベースを構築してモデルを評価する(ここで構築したのはテストコーパスであり，訓練データから分割されたのはテストセットである).

4.2.1 学習者コーパス

ICNALE（International Corpus Network of Asian Learners of English）はアジアの英語学習者の言語能力を研究するためのデータベースである．国際共同研究プロジェクトによって作成され，アジアの学習者からの英語の言語サンプルを収集し，分析することを目的としている[11]．ICNALE コーパスには，中国，日本，韓国など複数のアジアの国々の英語学習者達書いた英語エッセイが収集されている．これらのコーパスは，異なる年齢，学習バックグラウンド，英語レ

ベルを持つ学習者を代表している．本研究ではこのコーパスを用いて，学習者のエッセイを国別，英語レベル別に分類をやりました．

4.2.2 ニュースコーパス

ウィキペディアの記事は事実や情報を提供することを目的としている．個人的な意見や感情を含まず，中立的で客観的な立場から書かれている．記事の信頼性は信頼できる情報源や確立された知識に基づいていることが求められている．また，正確な情報を提供するためには，明瞭さと簡潔さの原則に従ってフォーマルで学術的なスタイルで書かれることが一般的である．ニュースの文では，報道機関やライターの視点を反映した評価やコメント，出来事の説明が含まれる．これは，報道機関やライターが自身の立場や意見を読者に伝えるための手段となりる．また，ニュース文はやすさを追求し，スラングや引用，読者の注意を引くための工夫がある．ウィキペディアの記事とニュースは全く異なる種類の文であるため，本研究ではニュース記事を収集し総数約 400 文のテストコーパスを構築して，さまざまな翻訳ツール(Google，Deepl，GPTv4)を使って英語に翻訳する．

4.2.3 論文データコーパス

論文には，ウィキペディアよりも深く掘り下げた信頼性の高い文章が含まれており，独創的な研究と研究テーマの深い分析が行われる．比較対象として，本研究では立命館大学の様々な分野の博士論文のアブストラクトから集めた日本語の論文を用いて日本語論文コーパスを構築した．また，電子情報通信学会で発表された日本人書いた英語論文を収集し，英語論文データコーパスを構築した．両方のコーパスにどちらも約 400 文が含まれている．

4.3 性能評価

モデルを評価するために，まず本研究では学習データから分割されたテストセットの分類結果適合率を計算した，図 13 から図 19 までは，各単語数英文異なる OOV 割合の再現率である，図 13，図 14 から分かるように，英文の単語数が 20 以下の場合は再現率がより低い，一方どんな長さの文でも基本的に日本人英語の再現率は 95%以上であり，単語数 20 以上の文では，ほとんどのネイティブ

英語の再現率は 90%以上である．結論として 20 語未満の文においては，ネイティブ英語と日本人英語の間にスタイルバイアスは見られない，20 語以上の文ではネイティブ英語と日本人英語にスタイルバイアスがあることが分かりました．

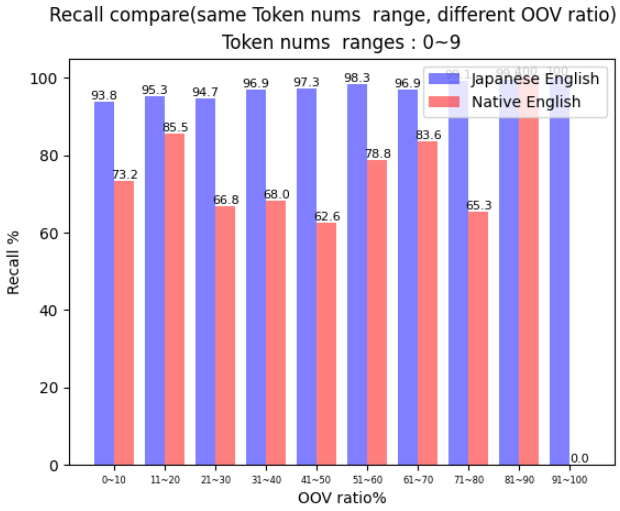

図 13. 文単語数 0~9 各 OOV 割合の再現率

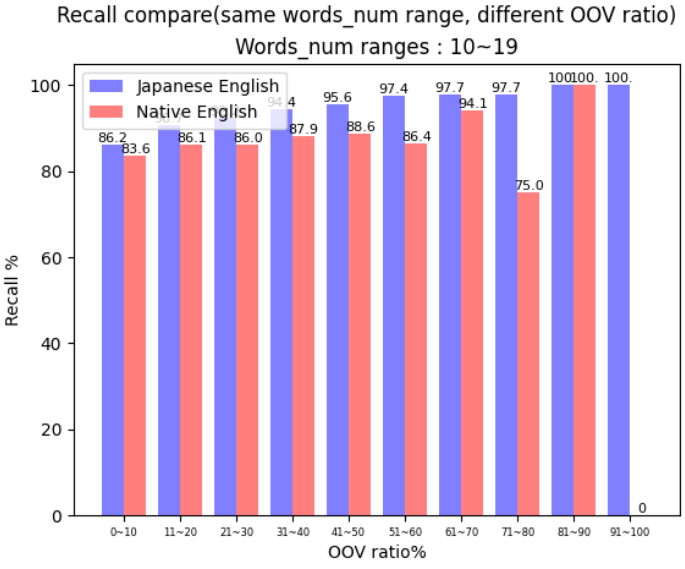

図 14. 文単語数 10~19 各 OOV 割合の再現率

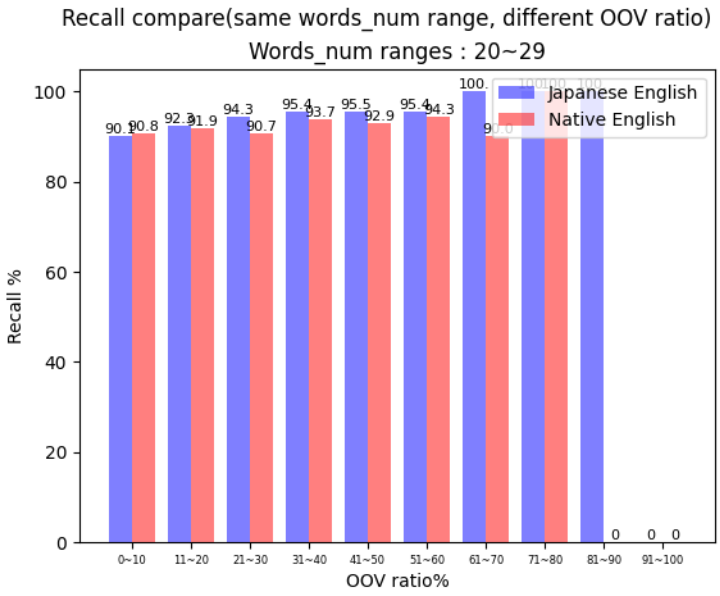

図 15 文単語数 20~29 各 OOV 割合の再現率

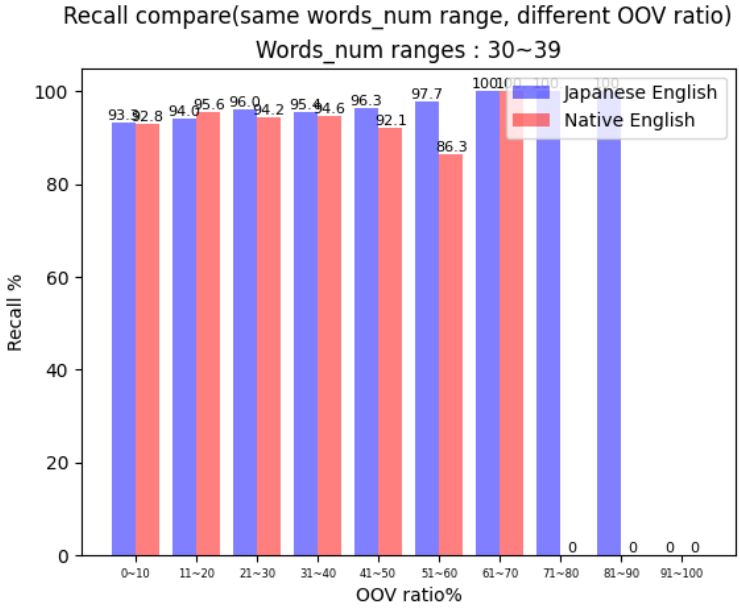

図 16 文単語数 30~39 各 OOV 割合の再現率

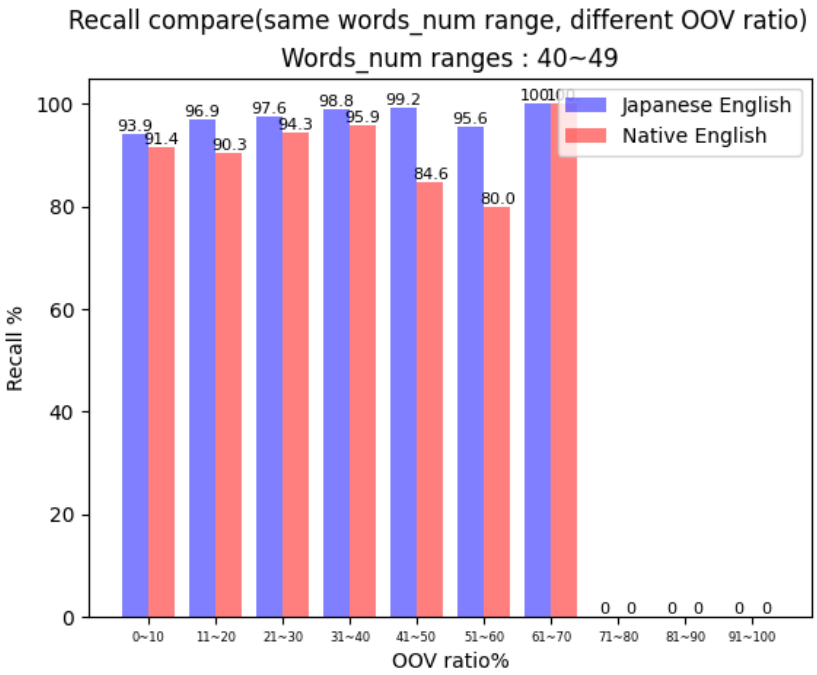

図 17 文単語数 40~49 各 OOV 割合の再現率

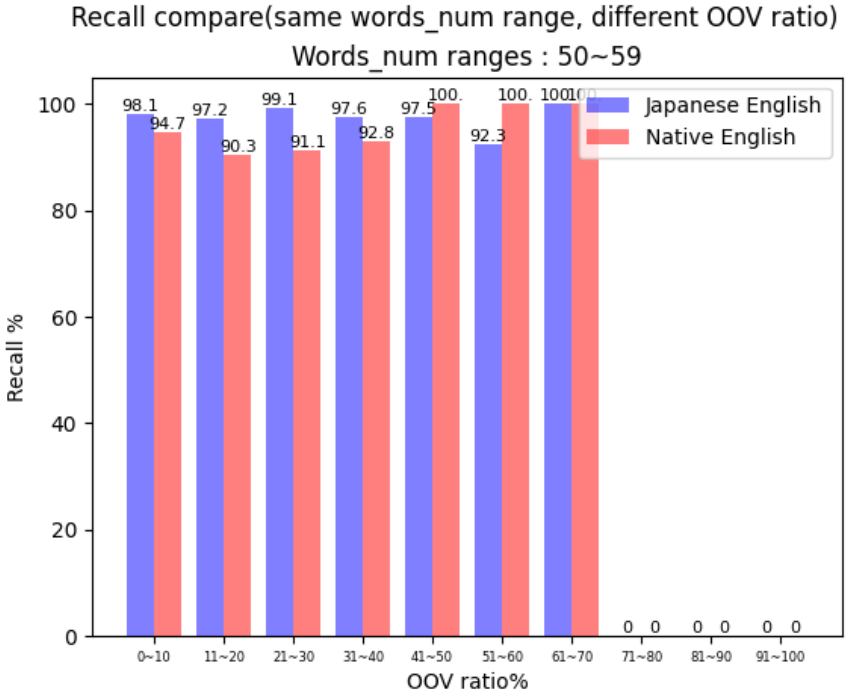

図 18 文単語数 50~59 各 OOV 割合の再現率

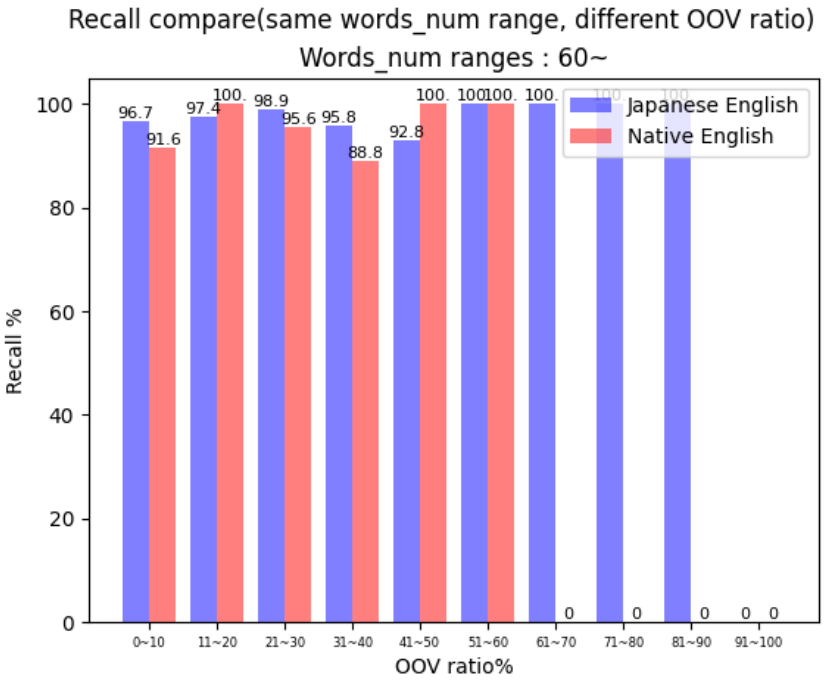

図 19 文単語数 60 以上各 OOV 割合の再現率

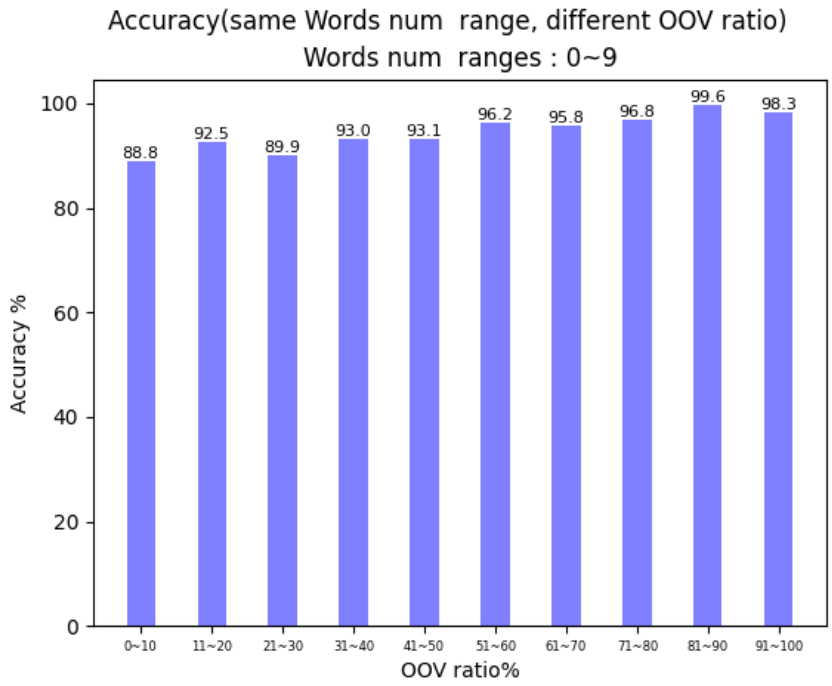

図 20. 文単語数 0~9 各 OOV 割合のアキュレシー

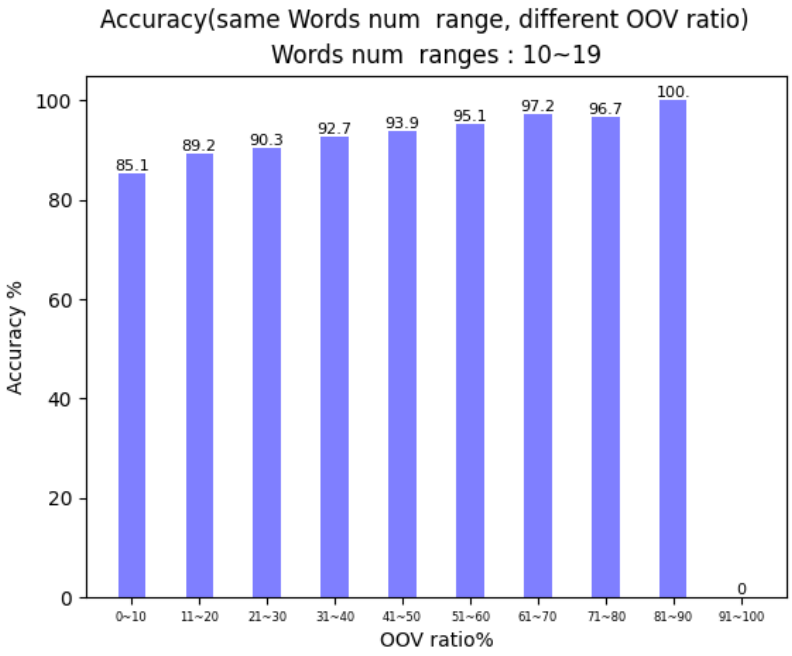

図 21. 文単語数 10~19 各 OOV 割合のアキュレシー

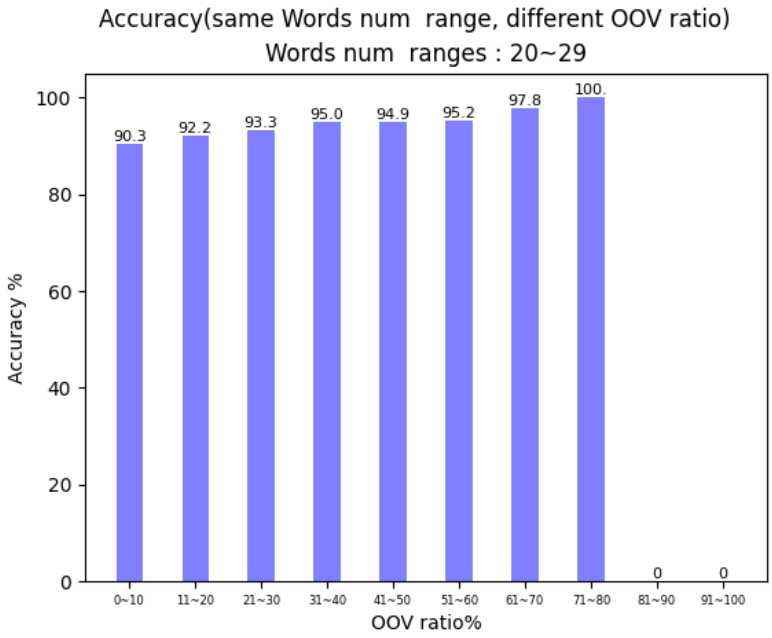

図 22. 文単語数 20~29 各 OOV 割合のアキュレシー

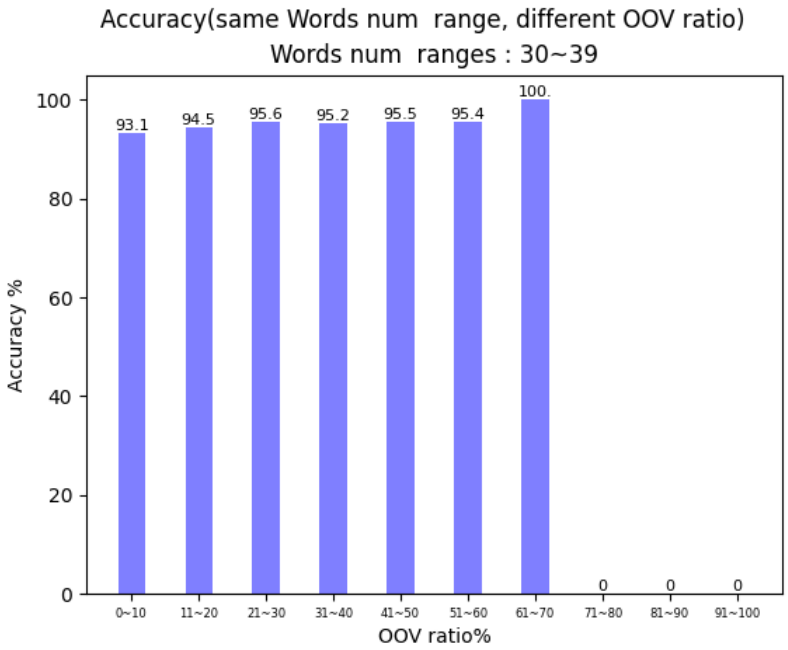

図 23. 文単語数 30~39 各 OOV 割合のアキュレシー

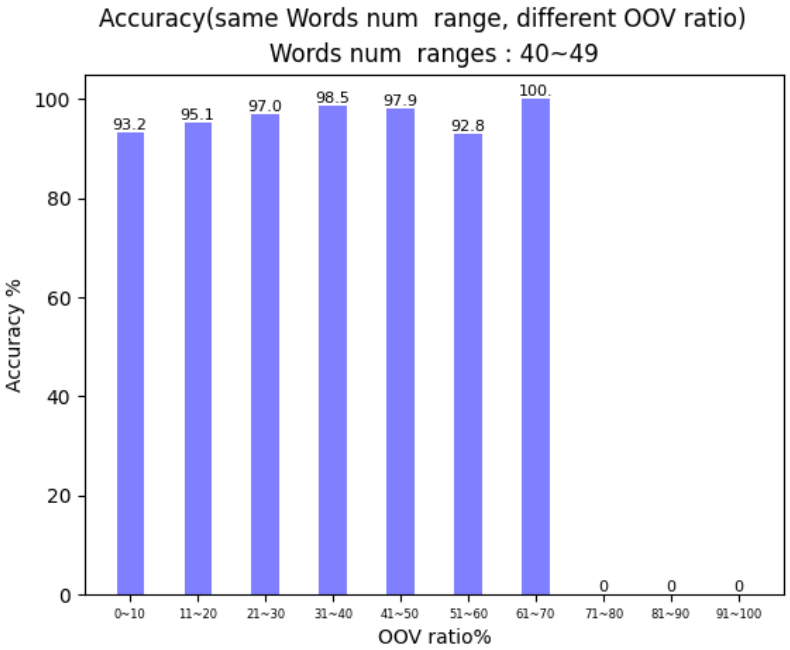

図 24. 文単語数 40~49 各 OOV 割合のアキュレシー

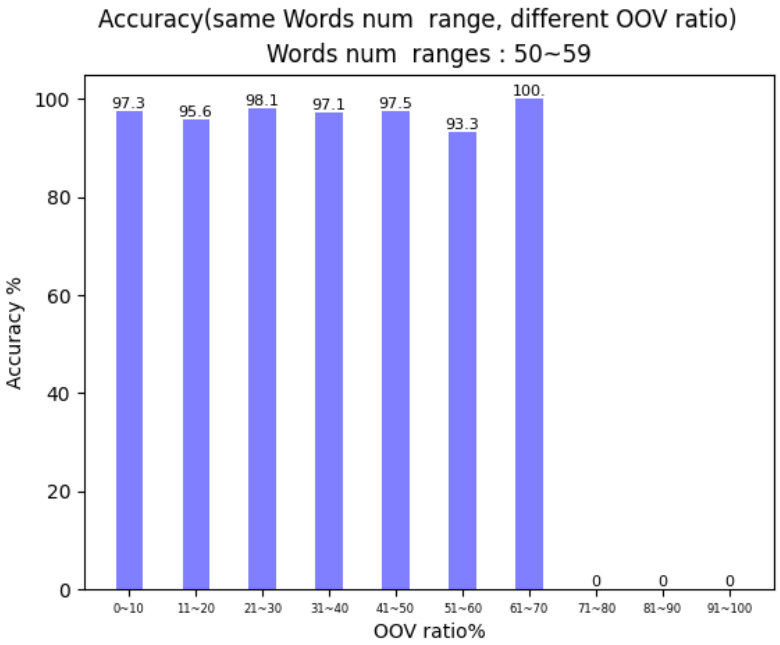

図 25. 文単語数 50~59 各 OOV 割合のアキュレシー

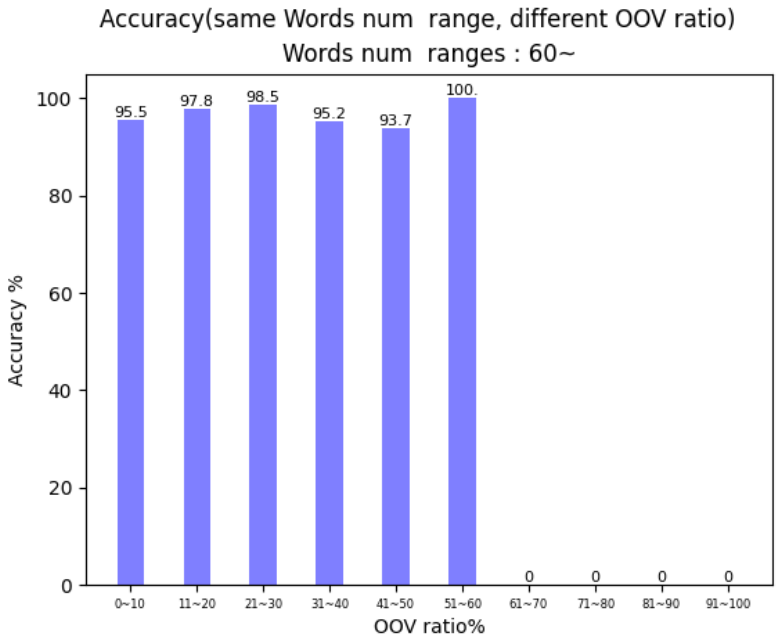

図 26. 文単語数 60 以上各 OOV 割合のアキュレシー

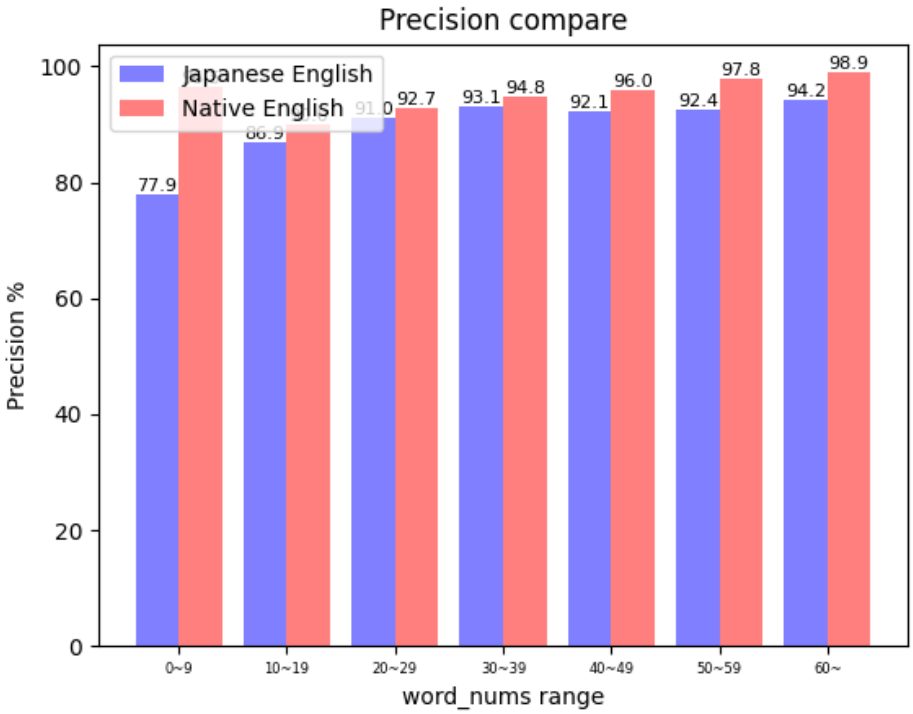

図 27. 各単語数範囲の適合率

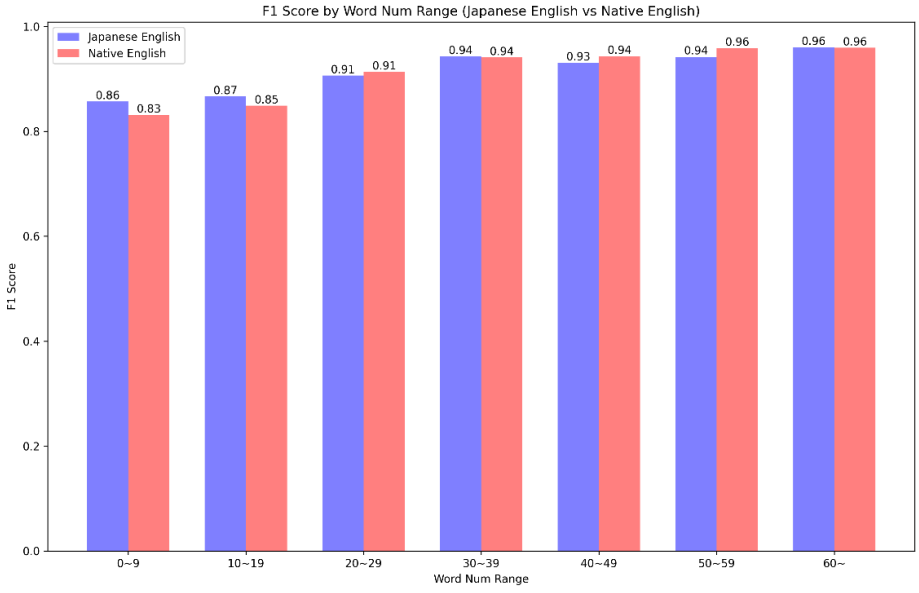

図 28. 分類器の F1 スコア

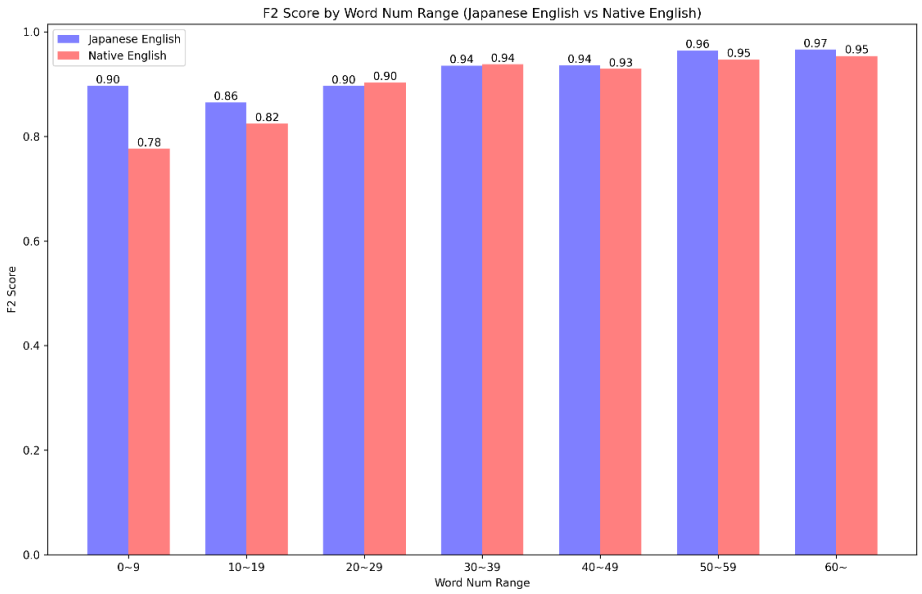

図 29. 分類器の F2 スコア

図 20~26 は，各単語数の範囲各 OOV 割合に対する分類器のアキュレシーを示している，単語数が 20 以下の文を分類するの結果と比べて，単語数が 20 以上だと，分類器のアキュレシーはより良いである．図 27 は分類器の適合率を示し，横軸は文の単語数を示す，OOV 割合ごとの Precision 本研究では計算してない，原因としては例えば，ある単語数範囲内，日本人英語が 50 文，ネイティブ英語が 500 文ある，両方の再現率を 90%とすると，この時日本人英語の Precision は 47.3%，ネイティブ英語の Precision は 98.9%であり，OOV の分布は分類器の評価に影響を与える可能性がある，そのため適合率，F1，F2 スコアを OOV 割合ごとの評価本研究では参考になれない．図 27 によると，日本語分類タスクでは，文の長さが 10 以内の場合，適合率は低いが，図 13 によれば，日本人英語分類タスクでは再現率が高く，それに対応して適合率も低くなるのは当然のことである．図 28 と図 29 は，分類モデルの F1 スコアと F2 スコアを示している．文章の長さが 20 以上の場合，モデルのスコアは 0.9 以上となっています，一方，20 以下の場合，スコアは 0.9 に達しないが，それでも良い性能が持っている．結論としては，このモデルは英語のスタイル分類タスクにおいて優れた能力を発揮しており，信頼性のある分類結果を提供している．

以上の結論を検証するために，本研究では，英語ネイティブが書いた SAT エッセイと ICNALE コーパスの日本人英語学習者全員が書いたエッセイを使って，

分類器で 2 つのコーパスを分類し，その結果を表 5 に示す．478 文のうち 73 文が日本人の英語と判定されました．その中で未分類区間に含まれる文が 8 文あるが，それらを分類失敗とみなすと，日本人の英語と判定された文は全体の 13.6%に過ぎない．一方，日本人学習者のエッセイのうち 88.5%が正しく分類されている．ここでの割合は再現率と同じ計算方法である．どちらのエッセイに対しても分類器の再現率は 85%以上である，上記の結論は正確である．

表 5. エッセイ分類結果

	日本人英語判定数	未分類区間数	総数	割合	未分類区間を除いた割合
SAT エッセイ	73	8	478	15.3%	13.6%
日本人学者エッセイ	11929	317	13123	90.9%	88.5%

ICNALE 学習者コーパスについて，本研究では学習者が書いた文を学習者の能力と母語でグループを分けてそれぞれ分類し,その結果を表 6 の通りに示す．レベル列は学習者の CEFR のレベルを示しており，A2_0 は TOEIC で 545 点以下に相当し，B1_1 は TOEIC で 550 点以上，B1_2 は TOEIC で 670 点以上，B2_0 は TOEIC で 785 点以上に相当する．表 6 からわかるように，学習者のレベルが上がるにつれて，日本人の英語と判定される割合が下がり．ENS とは英語ネイティブの結果です，SIN はシンガポール，どちらの割合も低いのは，どちらの英語もネイティブレベルであり，これは分類器が正しく分類したからである．従って，分類器の分類基準は学習者のレベル分けの基準と関連しているという仮設を提出する．関連研究を調べると，文長さに基づく複雑さの尺度は習熟度をよりよく予測する．語彙の複雑さについては，異なる単語の数の方が，習熟度によって書き手を区別できる[13]．異なるレベルの学習者の間には書いた文の複雑さに差があることが明らかになりった．そして，本研究は学習データから分割したテストセットと ICNALE コーパスにある日本人学習者のエッセイを用いて文の複雑さと分類器の分類基準との相関計算を行った．その結果は表 7,8 に示す．ほとんどすべての指標が分類確率に関連している．

表 6. 学習者エッセイ分類結果

国	レベル	文総数	日本人英語判定数	日本人英語判定数未分類区間	割合%	未分類区間を除いた割合%
CHN	A2_0	1406	1137	798	80.87	56.76
CHN	B1_1	6704	5157	3537	76.92	52.76
CHN	B1_2	3153	2358	1530	74.79	48.53
CHN	B2_0	376	261	165	69.41	43.88
ENS	XX1.	1802	804	386	44.62	21.42
ENS	XX2.	821	258	108	31.43	13.15
ENS	XX3.	1009	343	158	33.99	15.66
HKG	A2_0	32	17	9	53.13	28.13
HKG	B1_1	831	638	474	76.77	57.04
HKG	B1_2	1380	956	653	69.28	47.32
HKG	B2_0	448	289	187	64.51	41.74
IDN	A2_0	834	698	588	83.69	70.5
IDN	B1_1	2246	1998	1731	88.96	77.07
IDN	B1_2	2369	1886	1457	79.61	61.5
IDN	B2_0	68	38	25	55.88	36.76
JPN	A2_0	5066	4648	3676	91.75	72.56
JPN	B1_1	5974	5414	4213	90.63	70.52
JPN	B1_2	1522	1378	1048	90.54	68.86
JPN	B2_0	569	494	368	86.82	64.67
KOR	A2_0	2323	2041	1627	87.86	70.04
KOR	B1_1	2189	1947	1587	88.94	72.5
KOR	B1_2	2946	2470	1949	83.84	66.16
KOR	B2_0	2268	1830	1341	80.69	59.13
PAK	A2_0	615	552	468	89.76	76.1
PAK	B1_1	3018	2559	2063	84.79	68.36
PAK	B1_2	2793	2304	1856	82.49	66.45
PAK	B2_0	107	97	78	90.65	72.9
PHL	A2_0	44	22	12	50.	27.27
PHL	B1_1	248	174	128	70.16	51.61
PHL	B1_2	4407	2528	1549	57.36	35.15
PHL	B2_0	317	142	79	44.79	24.92
SIN	B1_2	3026	1048	519	34.63	17.15
SIN	B2_0	1492	380	164	25.47	10.99
THA	A2_0	3496	3065	2596	87.67	74.26
THA	B1_1	4963	4424	3678	89.14	74.11
THA	B1_2	2742	2377	1933	86.69	70.5
THA	B2_0	64	56	44	87.5	68.75
TWN	A2_0	783	682	519	87.1	66.28
TWN	B1_1	2380	2006	1536	84.29	64.54
TWN	B1_2	1633	1257	879	76.97	53.83
TWN	B2_0	572	417	288	72.9	50.35

表 7. ネイティブ英語に分類される確率と複雑さ指標とのスピアマン相関検定係数(テストセット)

Variable	Correlation (r)	p-value
単語数(W)	0.052906	0.0177
文数(S)	-0.03873	0.0826
動詞句 (VP)	0.01191	0.5937
節 (C)	-0.0656	0.0033
T ユニット(T)	0.000978	0.9651
従属節(DC)	-0.07527	0.0007
複合 T ユニット(CT)	-0.04308	0.0535
並列節(CP)	0.085829	0.0001
複合名詞(CN)	0.077414	0.0005
文の平均長さ(MLS)	0.060212	0.0069
T ユニットの平均長さ(MLT)	0.148919	1.97e-11
節の平均長さ(MLC)	0.257667	7.88e-32
文あたりの節数(C/S)	-0.02899	0.194
T ユニットあたりの動詞句数(VP/T)	0.078483	0.0004
T ユニットあたりの節数(C/T)	0.03871	0.0828
節あたりの従属節数(DC/C)	-0.0345	0.1221
T ユニットあたりの従属節数(DC/T)	-0.04107	0.0657
文あたりの T ユニット数(T/S)	0.011747	0.5987
複合的な T ユニットの割合(CT/T)	-0.00262	0.9066
T ユニットあたりの並列節数(CP/T)	0.127538	9.67E-09
節あたりの並列節数(CP/C)	0.149392	1.7E-11
T ユニットあたりの複合名詞数(CN/T)	0.148009	2.62E-11
節あたりの複合名詞数(CP/C)	0.237712	3.32E-27

表 8. ネイティブ英語に分類される確率と複雑さ指標とのスピアマン相関検定係数(ICNALE 日本人学習者エッセイ)

Variable	Correlation (r)	p-value
単語数(W)	0.184151	1.5E-100
文数(S)	0.024511	0.00496
動詞句 (VP)	0.112128	5.08E-38
節 (C)	0.131432	1.06E-51
T ユニット(T)	0.091994	4.33E-26
従属節(DC)	0.105917	4.35E-34
複合 T ユニット(CT)	0.085712	7.51E-23
並列節(CP)	0.067848	7E-15
複合名詞(CN)	0.155313	1.02E-71
文の平均長さ(MLS)	0.176728	1.25E-92
T ユニットの平均長さ(MLT)	0.143297	3.25E-61
節の平均長さ(MLC)	0.063377	3.58E-13
文あたりの節数(C/S)	0.122439	4.72E-45
T ユニットあたりの動詞句数(VP/T)	0.069528	1.49E-15
T ユニットあたりの節数(C/T)	0.087619	8.25E-24
節あたりの従属節数(DC/C)	0.079177	1.01E-19
T ユニットあたりの従属節数(DC/T)	0.089433	9.65E-25
文あたりの T ユニット数(T/S)	0.078739	1.6E-19
複合的な T ユニットの割合(CT/T)	0.063617	2.92E-13
T ユニットあたりの並列節数(CP/T)	0.056994	6.28E-11
節あたりの並列節数(CP/C)	0.032762	0.00017
T ユニットあたりの複合名詞数(CN/T)	0.132111	3.19E-52
節あたりの複合名詞数(CP/C)	0.092449	2.47E-26

本研究では NeoSca を使って文の複雑さ指標を計算した[14,15]，表 7,8 には，合計 23 の指標がある．最初の 9 つの指標は文構造の複雑性を表し，残りの 14 つの指標は文法の複雑性を表している，各指標の詳細は付録を参照する．結果としては，ほとんどの指標が分類時の確率と関連していることを示しており，つまり分類器の分類基準は文の複雑性に基づいて行われていると言えます．これらの指標の中で重みが高いのは T と CN/T である，どちらの相関係数も 0.1 を超

えた.

4.4 分類根拠の分析

本研究でトレーニングした英語スタイル分類器は，アテンション機構を活用した分類モデルであり，優れた性能を持つと言える．このモデルは，分類タスクにおいて高い精度を実現している．この高い精度の原因を分析するために，本研究はモデルの分類プロセスを可視化した.

4.4.1 アテンション可視化

分類器の性能は注意力機構が文の内容と構造を理解することから生じている．図 2 によると，ファインチューニングを行う際に，モデルは最終層に[CLS]というトークンの注意力を出力した，つまり，分類器は[CLS]の注意力値に従って分類をやる．[CLS]と文中の他の単語と比べて，この意味のないトークンはより公平に文中の各単語の意味情報を融合し，より良い文全体の意味を表している．具体的には，自己注意機構は文中の他の単語を利用してターゲット単語の意味表現を強化しますが，ターゲット単語自体の意味は依然として主要な要素となり，そのため，BERT の 12 層を経ることで，各単語の埋め込みはすべての単語の情報を融合し，自らの意味をより良く表現することが可能となる．一方，[CLS]は意味がない，12 層を経ることでアテンション後の全単語の加重平均を得るため，他のトークンと比べて，文の意味をより良く表現できる.

分類器結果のアテンションテンソルのサイズは[12，batch_size, num_heads, sequence_length, sequence_length]である，12 はモデルが 12 層のエンコーダを持ち，12 層の注意力があるといういみである，batch_size は，モデルが一度に処理できる文の数を表す，num_heads はマルチヘッド注意力を表す，sequence_length は文の長さを表す, [sequence_length, sequence_length]というのは単語が他のすべての単語に対して持つ注意力を表す. 可視化する際に, 12 個マルチヘッド注意力を全て足し合わせて使う，各層の注意力別々に表示されている. 図 30 は例である. トークンが赤いほど重みが高いである, 分類モデルは，主に中間層と最終層で情報を構築する[12]，第 5 層から 12 層まで見ると，モデルがアテンションしている単語はほぼ動詞と名詞である, 表 7,8 にも動詞と名詞に関連する指標すべて正の相関がある，例えば，VP/T，CN/T，CP/C．両者の結

論は一致している．

図 30．注意力可視化の例

第5章 スタイルバイアス分析

本章では，機械翻訳の翻訳結果を分析し，見つかったスタイルバイアスを紹介する.

5.1 機械翻訳

5.1.1 ニューラル機械翻訳

ニューラル機械翻訳は機械翻訳技術の一種であり，深層学習のニューラルネットワークを使用して自然言語間の翻訳を行う技術である．近年，NMT は大きな進化を遂げ,より長い文や複雑な文法構造に対して優れた性能を発揮し,言語間の翻訳の品質を向上させました.

現在翻訳ツールは多くの種類があり，最も有名なものは Google 翻訳，Deepl 翻訳, そして最近人気の GPTv4 である. これらの翻訳ツールの能力をテストし，翻訳結果のスタイルバイアスを分析するために，本研究ではこれらの翻訳ツールの結果を分析する.

翻訳結果を分析するために，いろいろの日本語原文を用意した，日本語 Wiki ページコーパス，モデルの訓練データと違って，まったく別のコーパスです ; 日本語ニュースコーパス，新聞サイトからコピーした記事のコーパス ; 日本語論文コーパス，立命館大学が公開した博士たちの論文のアブストのコーパスである. 3 つのコーパスを翻訳ツールで翻訳し，その結果を分析する.

Google 翻訳は，Google Neural Machine Translation（GNMT）という技術を使用して翻訳を行う．GNMT は 2016 年 11 月に発表されたニューラル機械翻訳システムであり，ニューラルネットワークを使用して Google 翻訳の流暢さと精度を向上させることを目指す. ニューラルネットワークは主に 2 つのモジュール，エンコーダとデコーダで構成されており, どちらも LSTM 構造を採用している. DeepL は Transformer を使用している．Transformer モデルの特徴は長い依存関係や文脈情報を同時に処理できることであり，優れた並行処理能力を持っているため，長い文の翻訳でも優れた性能を発揮する．DeepL はトレーニングプロセスで大量の対訳コーパスを使用し，異なる言語間のマッピング関係を深層学習によって学習し，高品質な翻訳結果を生成している．Deepl はイギリス英語とアメリカ英語の 2 種類の英語を生成することができる. 結果分析する際に，本

研究ではこの 2 種類の結果両方とも分析した．GPT（Generative Pre-trained Transformer）は，OpenAI が開発した自然言語処理のための深層学習モデルである．GPT は Transformer のデコーダをベースにしており，大量のデータを用いて事前学習された後，さまざまな自然言語処理のタスクに適応させることができる，それに，GPT は自己回帰モデルとして設計されており，テキストを生成する際には過去のコンテキストを利用して次の単語を予測する．これにより文の論理性や文法的な正確性が向上する，GPT-4 は OpenAI が新しく公開した，特大のマルチモーダルモデルであり，様々な自然言語処理のタスクにおいて驚異的な性能を示し，文章生成，質問応答，翻訳，要約などのさまざまなタスクに対応できる．

5.1.2 GPT による翻訳

GPT で翻訳のやり方は，GPT に指示を与える必要がある．本研究では以下の 2 種類のプロンプトを提示した，1 つ目は「You will be provided with a sentence in Japanese, and your task is to translate it into English.」，二つのは「You will be provided with a sentence in Japanese, and your task is to translate it into English as a native speaker.」，ネイティブとして翻訳することを強調しているかどうかの違いがある．その翻訳結果は同じではい，例えば，原文は「収穫直前の桃，およそ 1000 個が山梨県笛吹市の畑からなくなっていたことが警察への取材で分かりました.」，前者の結果は「It was found out through an interview with the police that about 1,000 peaches, just before harvest, had disappeared from a field in Fuefuki City, Yamanashi Prefecture.」，後者の結果は「It was discovered through police investigations that approximately 1,000 peaches, ready for harvest, had disappeared from a field in Fuefuki City, Yamanashi Prefecture.」．この 2 つの結果は，単語の選択が異なっている．

5.2 分析

表 9,10,11 に各コーパス翻訳文の分類結果を示す．各翻訳文の分類結果を見ると，Google 翻訳の翻訳結果が日本人英語に分類される割合が最も高いである．GPTv4 native の結果はネイティブ英語に最も近いである．つまり，GPTv4 native の翻訳結果は，単語が最も豊富で，文構造が最も複雑である．Deepl が提供する

英語の翻訳は，British English と America English の 2 つの異なるスタイルであり，前章の結論に基づいて，British English 翻訳文の方が America English よりも複雑さが高く，使った異なる単語の数も多いである．GPT-4 の場合，通常のプロンプトを提示しても，Google 翻訳や Deepl の結果よりもネイティブ英語に近いである，これがモデル性能の差別である．改善したプロンプトを使うと，GPTv4 のパフォーマンスは向上した，現在使用されているプロンプトはいずれも 1 文のシンプルなもので，今の結果は GPTv4 の上限ではい可能性がある．

表 9. Wiki コーパス翻訳結果の分類

Wiki コーパス	総数	日本人英語判定数	グレーゾーン数	割合	未分類区間を除いた割合
Google 翻訳	3000	1750	70	0.583	0.56
Deepl (America English)	3000	1474	65	0.491	0.469
Deepl(British English)	3000	1199	69	0.4	0.377
GPTv4	400	179	62	0.448	0.293
GPTv4native	400	162	58	0.405	0.26

表 10. ニュースコーパス翻訳結果の分類

ニュースコーパス	総数	日本人英語判定数	グレーゾーン数	割合	未分類区間を除いた割合
Google 翻訳	393	227	40	0.577	0.476
Deepl (America English)	392	146	67	0.372	0.201
Deepl(British English)	395	100	39	0.253	0.154
GPTv4	447	259	73	0.579	0.416
GPTv4native	447	240	83	0.537	0.351

表 11. 論文コーパス翻訳結果の分類

論文コーパス	総数	日本人英語判定数	グレーゾーン数	割合	未分類区間を除いた割合
Google 翻訳	319	172	38	0.539	0.42
Deepl (America English)	319	120	43	0.376	0.241
Deepl(British English)	319	110	43	0.345	0.21
GPTv4	319	168	56	0.527	0.351
GPTv4 native	319	161	52	0.505	0.341

5.3 検出されたスタイルバイアス

前章の結論によれば，文中の動詞や名詞が豊富であればあるほど，また文の構造が複雑であればあるほど，ネイティブ英語と分類される確率が高くなる．まず，「She started because she heard her uncle Kevin playing when she visited his house.」という文の分類結果は(native 英語 : 0.209,日本人英語 0.791)，「She was inspired to start because she heard her uncle Kevin playing when she visited his house.」になると，分類確率は(native 英語：0.902,日本人英語 0.098)になりました，“was inspired to”を使うと文複雑さいくつかの指標が上がりました，W，MLS，MLC が 14 から 17 になった，VP は 4 から 5 になった，MLC が 4.67 から 5.67 になった，相関係数から見ると，これらの指標の重みが高いである．それに，T ユニットあたりの複合名詞数（CN/T）という指標も重みが高いということが気づいた，この指標に関連する同格文という形式の文がある，同格文は複数の主語や述語がある形式の文である，文の主語は通常名詞である，CN/T と対応する．そこで，本研究ではいくつかの同格文を分類し，その結果を表 12 に示す．10 文中 7 文がネイティブ英語に分類される．W，MLS，MLC，VP，CN/T，これらの指標は分類器に大きな影響を与える．英語母語話者は通常より複雑な文を書き，文中に現れる単語の数も多いである．

表 12. 同格文の分類結果

同格文	ネイティブ英語確率	日本人英語確率
His passion, playing the guitar, is evident in every performance.	0.72	0.28
Our goal, to create a better world, guides our actions.	0.79	0.21
The belief that hard work leads to success is widely accepted.	0.52	0.48
The fact that she speaks five languages impressed everyone.	0.32	0.68
The rumor that they won the lottery turned out to be false.	0.17	0.83
His dream, to become a professional athlete, drives his dedication to training.	0.77	0.23
The idea that education is a lifelong journey is embraced by many.	0.74	0.26
The realization that time is precious changed her perspective on life.	0.86	0.14
The hope of finding a cure for the disease inspires scientists worldwide.	0.98	0.02
The discovery of a new species, the rare blue butterfly, fascinated researchers.	0.91	0.09

5.4 考察

機械翻訳の性能に着目すると，分類器の結果では，ICNALE コーパスに含まれる ENS や SIN の英語ネイティブが書いた英文が日本人英文に分類される確率は 20%～30%であり；機械翻訳の結果では，最も性能の高い GPTv4native が日本人英語に分類される確率は 30%～40%である．GPTv4native の性能は一般的な翻訳タスクに対応するのに十分であり，Google や Deepl の場合は結果を得た後，ユーザーが自分で使い勝手を判断する必要がある，全体から言うと機械翻訳の言語モデルの性能を向上させる余地が残されている．

第 4 章の結論では，分類器の分類基準は主に文の複雑さに基づいていることが示されている．ただし，他のいくつかの実験では，文の複雑さだけでなく，，語の選択や単語の順序が実験結果に影響を与える可能性もあるということが示唆されている．表 13 の示す通り，最初の 3 つの文では，単語の選択は同じで，文の意味も同じで；一番目の文の分類結果は大きな違いを示している．最後の 3 文では，“meal”が“dinner”に置き換るだけで，分類結果に大きな違いをもたらした．これらの操作は文の複雑さを変えることはないが，間違いなく分類器の分類結果に影響を与えた．この現象の原因は，分類器が英語を英語母語話者の単

語の使用習慣を学習していることと考えられる．そのため，分類器は判定の際に文中の単語の使用が英語母語話者の使用習慣に合っているかどうかを考慮する可能性がある．

表 13. 単語の選択や順番の実験結果

単語選択や選択異なる文	ネイティブ英語確率	日本人英語確率
After finishing his meal, John went for a walk in the park.	0.839619	0.160381
In the park, John went for a walk after finishing his meal.	0.479854	0.520146
John went for a walk after finishing his meal in the park.	0.449491	0.550509
After finishing his dinner, John went for a walk in the park.	0.372732	0.627268
In the park, John went for a walk after finishing his dinner.	0.430854	0.569146
John went for a walk after finishing his dinner in the park.	0.015795	0.984205

第6章 おわりに

本研究では，ネイティブ英語と日本人英語のスタイルバイアスを検出するために BERT を用いた英語スタイル分類器をトレーニングし，バイアス検出手法を提案した．モデルをトレーニングするために，日本人英語コーパスとネイティブ英語コーパスを構築するした．モデルの性能をテストするために，ニュースコーパスと論文コーパスを構築した，構築したコーパスを用いて，分類器の分類確率と文の各文複雑さ指標の相関係数を算出し，分類器の分類基準と文の複雑さが関連していることが検証した．機械翻訳の性能も測った，GPTv4native は最高の性能が持っている．本研究の貢献は以下の通りである．

英語スタイルバイアス分類器の構築

日本人英語とネイティブ英語のデータベースを構築した．このデータベースには 60 万以上の英語文が含まれており，日本人英語とネイティブ英語の分類器を訓練しました．分類器の正解率は 90%以上に達した，テストコーパスとして，ニュースコーパスと論文コーパスを構築した，日本人英語とネイティブ英語の間に明確なスタイルバイアスがあることが証明した．

スタイルバイアスの分析

分類器の分類基準と文複雑さの指標の相関係数を算出し，分類器の出力と文複雑さ指標の間に関連性があるということが証明した．分類器の分類プロセスを可視化した．日本人英語とネイティブ英語の間のスタイルの違いを検出した．5 つの機械翻訳ツールの中で，Google 翻訳の結果は学習者の文に最も似ている，GPTv4native は最高の性能が持っており，Deepl(British)は Deepl(America)よりもネイティブに近いである．

本研究では，スタイルバイアスに関する分析は徹底していない．今後の展望として，ネイティブ英語コーパスを拡張し，短文分類の再現率を向上する．現時点で分類器の分類基準と文複雑さの間の相関係数を計算しましたが，表 7 と表 8 に示すように，相関係数にはまだ若干の差がある．この差は ICNALE 学習者コーパスのデータは偏りすぎって，つまり非ネイティブ文の割合が高すぎることに起因している可能性がある．他のコーパスを使って相関係数を算出した後，比較を行う必要がある．

また現時点で分類器の分類基準と文複雑さの間の相関係数 p 値はほぼすべてが 0.05 以下であることを示している．これにより，将来文の複雑さと分類確率

の関係を予測するためのモデルを訓練することが期待している．このモデルは 2 次元の予測確率を表す y と 23 次元の複雑さ指標を表す x を入力とする関数のようなものである．異なる複雑さ指標を入力することで，異なるタイプの文の分類確率を得ることができる，これにより，複雑さ指標以外の要因を分析するのに役に立つと考えられる．

最後に，文のスタイルに影響を与えるのは文の複雑さだけじゃなく，表 13 に示すよう単語の選択や順番など他の要素も文のスタイルに影響を与える．これらの要因については研究を進めるために詳細な調査する必要がある．

謝辞

本研究を行うにあたり，熱心なご指導，ご助言を賜りました指導教官の村上陽平教授に深謝申し上げます.また，この四年間普段からお世話になっている社会知能研究室の皆さまに心より感謝申し上げます.

参考文献

[1] Czopp A M, Kay A C, Cheryan S. Positive stereotypes are pervasive and powerful[J]. Perspectives on Psychological Science, 2015, 10(4): 451-463.

[2] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[3] May C, Wang A, Bordia S, et al. On measuring social biases in sentence encoders[J]. arXiv preprint arXiv:1903.10561, 2019.

[4] Xu H, Liu B, Shu L, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis[J]. arXiv preprint arXiv:1904.02232, 2019.

[5] Costa-jussà M R, de Jorge A. Fine-tuning neural machine translation on gender-balanced datasets[C]//Proceedings of the Second Workshop on Gender Bias in Natural Language Processing. 2020: 26-34.

[6] Biber D, Finegan E. Drift and the evolution of English style: A history of three genres[J]. Language, 1989: 487-517.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[8] Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.

[9] 独立行政法人情報通信研究機構，日英京都関連文書対訳コーパス，0113,2011

[10] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[11] Ishikawa S. Aim of the ICNALE GRA project: Global collaboration to collect ratings of asian learners' l2 english essays and speeches from an ELF perspective[J]. Learner Corpus Studies in Asia and the World, 2020, 5: 121-144.

[12] Peters M E, Ruder S, Smith N A. To tune or not to tune? adapting pretrained representations to diverse tasks[J]. arXiv preprint arXiv:1903.05987, 2019.

[13] Avci A.CAF across proficiency levels and profiles: an investigation of ESL student writings in an English placement test[D]. 2020.

[14] Long Tan. NeoSCA: A Rewrite of L2 Syntactic Complexity Analyzer, version 0.0.43.2022.

[15] Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474-496.

付録

表 14. 文複雑さ指標説明

Variable	Explanation
単語数(W)	単語の数
文数(S)	文の数
動詞句 (VP)	主要な動詞(不定形、現在分詞、過去分詞)とその修飾語(副詞、前置詞句など)からなる句の構造である. 動詞句は動作や状態を表現するために使用される.
節 (C)	節とは文の一部であり, 主語と述語で構成される比較的小さい言語単位である.
T ユニット(T)	文で分解できる文法的に完全な最小単位(最小終端可能単位), 主節とその主節に従属する従属節からなるの単位.
従属節(DC)	従属節とは独立して文を成すことができない節のことである, 主文の存在に依存し, 主文の意味を修飾または制限するために使われます.
複合 T ユニット(CT)	複雑 T 単位とは主文と 1 つまたは複数の従属節を含む T 単位のことを指す. 例えば形容詞節や副詞節などがある.
並列節(CP)	並列句とは 2 つ以上の同等な要素(単語, 文, または節)を並列接続詞で結んだ構造のことを指す. 並列句は文中で並列関係を表現するために使われます.
複合名詞(CN)	複雑名詞句とは中心となる名詞に加えて, 形容詞, 冠詞, 代名詞などの修飾要素を含む名詞句の一種である. 複雑名詞は名詞の意味を説明したり制限したりするために使用される.
文の平均長さ(MLS)	文の平均長さである, 単文の場合 W と同じ.
T ユニットの平均長さ(MLT)	T ユニットの長さ割る T ユニットの数.
節の平均長さ(MLC)	節の長さ割る節の数.
文あたりの節数(C/S)	文中に節の数.
T ユニットあたりの動詞句数(VP/T)	各 T ユニットの動詞句の数.
T ユニットあたりの節数(C/T)	各 T ユニット内の節の数.
節あたりの従属節数(DC/C)	各節に含まれる従属節の数.

T ユニットあたりの従属節数（DC/T）	T ユニットあたりの従属節の数.
文あたりの T ユニット数（T/S）	並列節の比率, 各文中の T ユニットの数.
複合的な T ユニットの割合（CT/T）	複合 T ユニット比率, 各 T 単位中の複合 T ユニットの数.
T ユニットあたりの並列節数（CP/T）	T ユニットあたりの並列節の数.
節あたりの並列節数（CP/C）	節ごとの並列節の数
T ユニットあたりの複合名詞数（CN/T）	T ユニットあたりの複合名詞文の数
節あたりの複合名詞数（CP/C）	各節の複合名詞文の数