ロングテイルな分布の入力を扱う機械学習システムに対するテスト設計手法の提案

研究員：後藤　優斗（コニカミノルタ株式会社）
　　冨井　良平（株式会社東光高岳）
　　松尾　正裕（パナソニック ITS 株式会社）
主査　：石川　冬樹（国立情報学研究所）
副主査：栗田　太郎（ソニー株式会社）
　　徳本　晋（富士通株式会社）

研究概要

本研究では，機械学習のデータセットによくみられる，大量のレアケースを含むロングテイルな分布において，機械学習の外部品質特性である AI パフォーマンスとリスク回避性を両立した品質評価手法について提案する．機械学習システムにおいて，複数の品質特性が挙げられるが，これらは必ずしも両立するものではなく，時には相反する特性となることもある．AI-OCR を題材とし，文字の使用頻度に着目してこれら 2 つの品質特性のバランスを考慮して評価する方法とその効果，また，そのときのテスト削減効果が十分有効であったため，その内容について報告する．

1.　はじめに

近年，システムやソフトウェアに機械学習を利用することが増えており，機械学習システムに対するテスト手法の研究や，品質保証のためのガイドラインが制定されている．しかし,機械学習システムは学習するデータセットによって振る舞いが決定づけられるため,画一的な品質保証の議論は難しく，ドメイン知識や経験に頼らざるを得ない部分がある．また，機械学習システムの品質には複数の要素があり，例えば，機械学習システムにおける外部品質特性として，AI パフォーマンス（有用性）やリスク回避性，公平性，プライバシー，AI セキュリティが挙げられる[1]．これらの品質特性は必ずしも両立できるものではなく，時には相反することもあるため，機械学習システムの品質は必ずしも予測精度のみで評価できるものではなく，使用される場面に応じて適切に品質評価を行う必要がある．

本研究では，機械学習システムの１つである，AI-OCR の品質評価を対象に，機械学習システムの品質保証における課題の解決方法について検討する．AI-OCR とは，機械学習を用いて画像から文字を認識し,文字毎に文字コードへ変換する技術[2]のことである.例えば，文字の種類が少ない英語を認識する AI-OCR の品質保証を考えると，アルファベット 52 種類（大文字 26 種類，小文字 26 種類）を正確に文字コードへ変換できるかを評価すればよい．しかし，日本語の AI-OCR の場合，評価対象となる文字の種類はアルファベットに加え，ひらがな，カタカナ，漢字が評価の対象となり，文字の種類を合計すると約 6 万文字に達する[3]．したがって，AI-OCR を愚直にテストしようとすると，用意すべきデータは膨大な数になる．一方，文化庁の漢字出現頻度数調査によると，常用漢字の 2,136 文字が一般の社会生活における漢字の使用状況において 98%を占めていると報告されている[4]．この調査結果を踏まえると，実際の利用場面を想定した精度評価では，常用漢字を中心にテストすることで，効率的なテストを行うことが可能であると考えられる．このような状況を想定して，本論文では，実際の利用時の精度を予測するため，AI-OCR の運用時における文字の使用頻度を考慮したテスト設計手法について検討した．

本論文の構成は以下の通りである．2 章では前述の課題認識に関して，本研究の背景や

動機についてまとめる．その点を踏まえて，3 章，4 章では今回の研究課題と提案手法について述べる．その後，5 章，6 章では AI-OCR を用いた精度評価に関する実験の内容と結果について報告する．最後に 7 章，8 章では実験結果に対する考察と今後の展望について議論する．

2.　背景・動機

AI-OCR の品質への要求として，実際の読み取り対象の文字をなるべく正確に読み取ること，また，より多くの種類の文字を幅広く認識できることの 2 つが挙げられる．前者は読み取り精度であり，AI パフォーマンスに該当する．ここで，読み取り精度とは，画像から文字を認識し，文字コードに正しく変換できた割合のことである．また，後者は対応していない文字があることにより AI-OCR への信頼を棄損する可能性が考えられるため，リスク回避性に該当する．AI-OCR において，これらの品質特性はトレードオフの関係になっているので，これらの品質特性のバランスを考慮して品質評価する必要がある．

また，AI-OCR におけるテストデータに対する要求として，データ件数の削減が挙げられる．機械学習システムに対する入力データは，原理的には無限に存在するため，効率の観点から，少ないテストデータで評価することが要求される．ここで，評価コストについて考えてみると，多くの AI システムにおける評価では収集したデータを用いるため，個々のデータに対して正解をラベリングするためのコストがかかる．一方，AI-OCR のテストデータは，任意の文字に対してテストデータを作成するため，テスト対象の文字を絞ることにより，テストの量を減らすことができる．そこで，今回は，文字の使用頻度の分布がロングテイルになっていることに着目する．使用頻度により，使用頻度の高い文字（以下，メジャーな文字）と使用頻度の低い文字（以下，マイナーな文字）に分類して考える．メジャーな文字は,使用頻度の高さと文字の種類が限定されることから,削減する必要はない．一方，マイナーな文字は，使用頻度の低さと文字の種類の多さから，すべての文字をテストすると費用対効果が悪いため，テスト対象となる文字の種類を絞り込む必要がある．

なお，本研究では，文字の使用頻度分布に着目して議論するため，通常の AI-OCR 評価で検討する文字の修飾や画像化における外乱等についての影響は無視することとする．

3.　研究課題

本研究では，前述した品質評価における観点について，3 つの課題としてまとめ，また，検証すべき内容として Research Question（以下，RQ）を定義する．各 RQ について実験を行い，複数のモデルを用意して比較検討することにより，有効性の検証を行う．

3.1.　AI パフォーマンス（有用性）の測定

ある会社では，AI-OCR における利用時精度は，常用漢字や JIS 第一水準漢字のような，使用頻度の高いメジャーな文字のみを評価することで測定している．ただし，実際に使用される文字の中には，マイナーな文字も一定数含まれているため，メジャーな文字のみを用いた精度測定により，AI パフォーマンスをどの程度正確に測定できるかを検証する．

3.2.　AI パフォーマンス（有用性）とリスク回避性の品質評価バランス

AI-OCR において，利用時精度の評価だけではなく，さまざまな種類の文字を読み取れることも確認する必要がある．これらをバランスよく評価する方法について検討する．次章にて，AI パフォーマンスとリスク回避性のバランスをとった評価手法を提案し，実験を行うことで手法の有効性について確認する．

3.3.　マイナーな文字の削減量

2 章で述べた通り，テスト対象の文字を削減するには，マイナーな文字を削減するほうが合理的である．3.2.節で述べた提案手法の有効性を確認する実験の中で，マイナーな文字をどの程度削減できるかを確認する．

3.4.　Research Question

本研究において検証する内容について，RQ としてまとめる．なお，同じ実験で検証すべ

き RQ も存在するため，それぞれの検証内容の関係性が分かるように採番する．
RQ.1-1　メジャーな文字のみを用いて評価することにより，利用時精度を予測することができること
RQ.2-1　提案手法を用いて評価することにより，AI パフォーマンスとリスク回避性のバランスが取れた評価を行うことができること
RQ.2-2　提案手法を用いて評価することにより，テストデータの数を削減できること

4.　提案手法

3.2.節で述べた内容を実現するために，メジャーな文字のみを用いる利用時精度の評価と，すべての文字が読み取れるリスク回避性の 2 つの特性のバランスを考慮した評価手法を検討する．また，AI-OCR で扱うデータの特徴として，文字の使用頻度がロングテイルな分布になっていることに着目して，以下に示す手法を提案する．

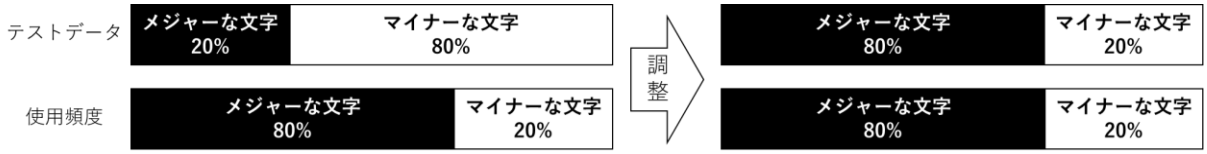

図 1　提案手法によるテストデータの作成イメージ

提案手法のイメージを図 1 に示す．はじめに，すべての文字を 1 回ずつ評価するようなテストデータを作成する．このとき，パレートの法則によると，メジャーな文字がテストデータの 20%を占め，マイナーな文字が 80%を占めることになる．一方，文字の使用頻度の累積度数に着目すると，メジャーな文字が 80%，マイナーな文字が 20%を占めることになる．この状態のテストデータを用いて精度評価を行うと，利用時精度とかけ離れた精度になることが考えられる．そこで，テストデータに占めるマイナーな文字の割合を 20%になるように，マイナーな文字を大幅に削減する．このとき，マイナーな文字の削減対象は，ランダムに選択する手法を用いて決定する．この操作を行うことにより，文字の利用頻度の累積度数におけるメジャーな文字とマイナーな文字の割合と，テストデータにおけるメジャーな文字とマイナーな文字の割合を近づけることができる．

実際の閾値は 80%が最適であるとは限らないため，任意の閾値に対して，この手法による文字の数を以下のように決定する．すべての文字数（種類）を n_{all}，メジャーな文字の文字数を n_{major}，マイナーな文字の文字数を n_{minor} とする．メジャーな文字を使用頻度 t % $(0 \leq t \leq 100)$ を閾値とすると，それぞれの文字数の関係は以下の通りになる．

$$n_{\mathrm{all}} = n_{\mathrm{major}} + n_{\mathrm{minor}} \tag{1}$$

$$n_{\mathrm{major}} = \frac{t}{100} n_{\mathrm{all}} \tag{2}$$

$$n_{\mathrm{minor}} = \frac{100 - t}{100} n_{\mathrm{all}} \tag{3}$$

$$n_{\mathrm{minor}} = \frac{100 - t}{t} n_{\mathrm{major}} \tag{4}$$

実際の品質評価時には，JIS 第一水準漢字などの規格を用いることを考慮すると，メジャーな文字は既知であるため，マイナーな文字の削減量のみを決定したい場合が想定される．この場合，(4)式により，削減後のマイナーな文字の文字数を求めることができる．

5.　実験

5.1.　実験における検証内容

本研究では，2 つの実験を行うことで，3 章で挙げた RQ の検証を行う．実験 1 では，3.1.節で述べた AI パフォーマンスの測定について検証する．実験 2 では，提案手法による有効性を確認することで，3.2.節と 3.3.節で述べた AI パフォーマンスとリスク回避性のバ

ランスを考慮した品質評価と，マイナーな文字の削減量について検証する．

5.2.　実験で使用する文字の頻度分布

本実験では，名字由来 net にある全国名字ランキングにおける上位 40,000 件の名字[5]を利用時データとして，それぞれ 1 回ずつ読み取ることを想定する．この名字 40,000 件に対して，使用されている文字の種類を一覧化し，使用回数をカウントすることで使用頻度分布を作成した．また，文字の種類について集計・整理したところ，実際に名字 40,000 件で使用される文字の総数は 85,070 字であり，文字の種類（文字数）は 2,245 字であった．作成した使用頻度の分布と主要な使用頻度における値を図 2，表 1 に示す．文字の使用頻度分布は，本実験で想定しているロングテイルな分布であることが確認できた．

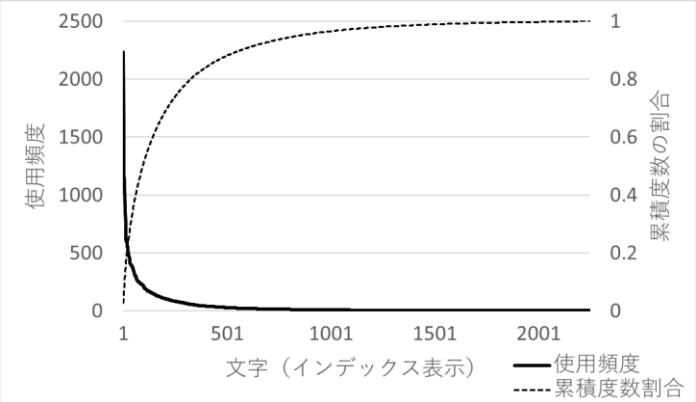

図 2　名字 40,000 件における文字の使用頻度分布

表 1　図 2 における主要な値

累積度数の割合	文字数	全文字数との比*
100	2,245	1.0000
95	875	0.3898
80	325	0.1448
60	143	0.0637

* 全文字数との比=文字数/全文字数

5.3.　AI-OCR モデルの作成

本実験では AI-OCR エンジンである Tesseract[6]を用いて，デフォルトで提供されているモデルと，特定の文字の読み取りに特化して学習したモデルを用いて評価を行う．本実験において比較評価するモデルは以下の 4 つである．なお，学習した文字数の順にモデルを並べているが，この順番で性能が良いという意味ではない．

- AI-OCR_Model_1: Tesseract に準備されている既存のモデル
- AI-OCR_Model_2: 使用頻度 100%までの文字（2,245 字）を学習したモデル
- AI-OCR_Model_3: 使用頻度 95%までの文字（875 字）を学習したモデル
- AI-OCR_Model_4: 使用頻度 80%までの文字（325 字）を学習したモデル

5.4.　実験内容

本研究では，3 章で定義した RQ について検証するため，2 つの実験を行い，3 つの内容に対して検証を行う．実験で使用する精度と実験内容についてそれぞれ述べる．

5.4.1.　それぞれのモデルにおける読み取り精度の測定方法

今回の実験における読み取り精度は，テスト対象の文字を 1 文字ずつ評価して，正しく読み取れた文字の割合である．テスト対象の文字を 1 文字ずつ評価するため，1 文字のみ記載された評価用画像をテスト対象の文字数分だけ用意して，AI-OCR により読み取れたかどうかを確認する．テスト対象の文字はメジャーな文字の閾値により決定されるため，読み取り精度は，メジャーな文字の閾値に依存して変化する．また，提案手法では，読み取り精度の評価にマイナーな文字を含めることになるが，計算方法は同じである．

また，利用時精度は，すべての文字を対象として，正しく読みとれた文字に対して使用頻度で重みづけを行い，計算した割合である．読み取り精度と同様に，テスト対象の文字を 1 文字ずつ評価し，読み取り結果を確認する．また，すべての文字を評価して利用時精度を計算するため，利用時精度はメジャーな文字の閾値によらず一定値になる．

以上の説明を数式としてまとめると，メジャーな文字の使用頻度 t％ $(0 \leq t \leq 100)$ を閾値としたときの読み取り精度 p_t と利用時精度 P は以下のように計算できる．ここで，文字 i の読み取り結果を r_i（正しく読み取れたら $r_i = 1$，それ以外は $r_i = 0$），評価対象の文字の集合を $Char_t$，評価対象の文字数を n_t，全文字数を N，文字 i の使用頻度を $f(i)$ とする．

$$p_t = \frac{1}{n_t} \sum_{i \in Char_t} r_i \tag{5}$$

$$P = \frac{1}{N} \sum_{i} r_i \times f(i) \tag{6}$$

5.4.2.　実験 1：メジャーな文字のみを用いた利用時精度の評価

実験 1 では，メジャーな文字のみを用いて評価した読み取り精度を使用して，利用時精度の評価が行えるかどうかを確認する．まず，(5)式を用いて，メジャーな文字の閾値 t を変化させながら，読み取り精度を算出する．その後，(6)式によりあらかじめ算出しておいた利用時精度を用いて，読み取り精度との乖離を確認する．また，読み取り精度が利用時精度と一致する点があるかどうかを確認する．したがって，5.3.で定義した AI-OCR_Model_1 から AI-OCR_Model_4 の 4 つのモデルに対してそれぞれ前述の精度計算を行い，次の 2 つの内容を確認することで，利用時精度を予測できるかどうかを検証する．

① それぞれのモデルに対して，読み取り精度と利用時精度の乖離状況
② 4 つのモデル間において，利用時精度と読み取り精度の順位が同じであること

5.4.3.　実験 2：提案手法による精度評価とテスト対象となる文字の削減

提案手法を用いてモデルの精度評価を行うことにより，AI パフォーマンスとリスク回避性のバランスを取れた評価が可能であることを確認する．実験 2 では，利用時精度は使用せず，モデル間における読み取り精度の順位について確認する．また，実験 1 と同様に，閾値 t を変化させながら，各モデルにおける精度の変化についても確認する．

更に，提案手法では，テスト対象に含めるマイナーな文字を大幅削減することで，テストデータの削減効果も見込んでいる．削減対象とされたマイナーな文字の数を確認することにより，テストデータの削減効果についても確認する．

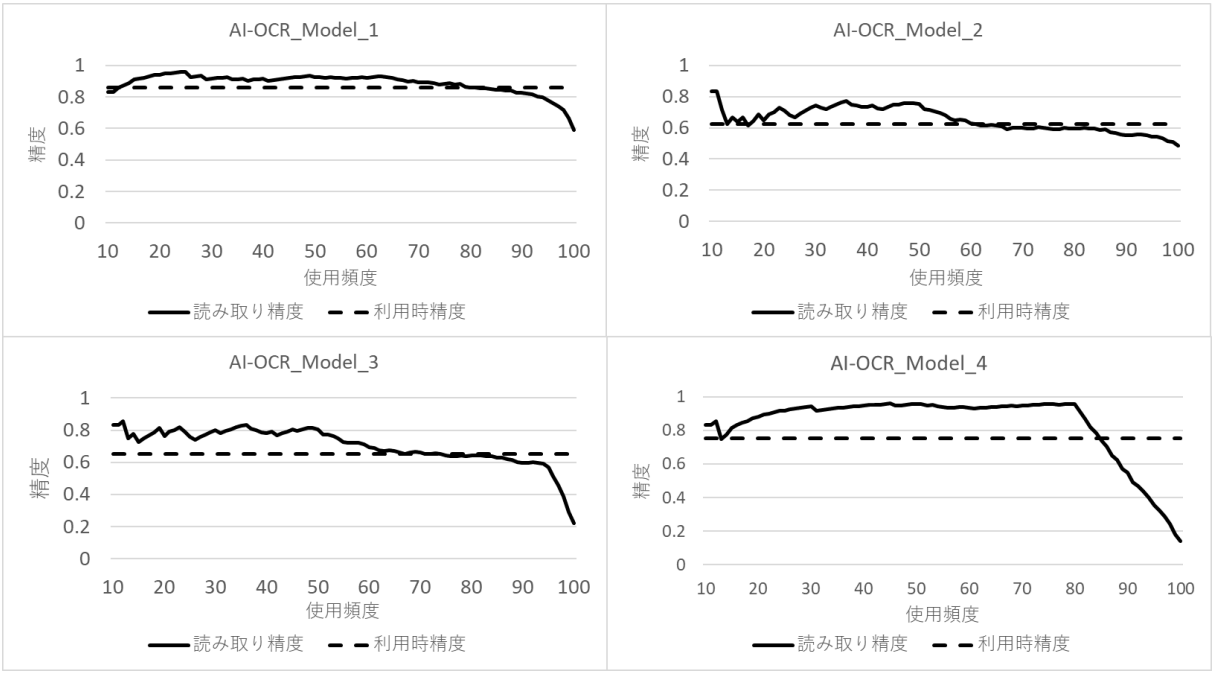

図 3　読み取り精度の評価結果と利用時精度

6.　実験結果

6.1.　実験 1:メジャーな文字のみを用いた精度評価

メジャーな文字のみを 1 回ずつ評価した時の読み取り精度と利用時精度のグラフを図 3 に示す．利用時精度はすべての文字を評価しているため，使用頻度によらず一定である．一方，読み取り精度は，メジャーな文字の閾値により評価対象の文字が変化するため，使

用頻度に依存して変化する．また，AI-OCR_Model_3 と AI-OCR_Model_4 は学習した文字の使用頻度を境に読み取り精度が急落している．これは，境界値以降の使用頻度では読み取れない文字のみが存在するため，使用頻度が増加することにより，読み取り対象の文字数のみが増加し，読み取れた文字の数は一定になっていることが要因である．

まず，AI-OCR_Model_1 から AI-OCR_Model_3 の各グラフに着目すると，メジャーな文字を十分に評価している使用頻度の範囲（60%から 80%）では，利用時精度と読み取り精度の差は小さく，2 つのグラフの交点もこの範囲に存在している．したがって，メジャーな文字のみを評価することで,利用時精度を予測することが可能であると考えられる.ただし，AI-OCR_Model_4 に関しては，同じ範囲において，読み取り精度が利用時精度よりも全体的に大幅に上振れしている．したがって，メジャーな文字のみを用いて評価する方法は，すべてのモデルに対して利用時精度を精度よく予測できる万能な方法とは言えない．

続いて，利用時精度と読み取り精度の順位を比較する．同じ使用頻度の範囲において，利用時精度が最も高いモデルは AI-OCR_Model_1 である．一方，読み取り精度が最も高いモデルは AI-OCR_Model_4 になる．それ以外のモデルにおける各精度の順位については同じである．したがって，メジャーな文字のみを用いて評価した場合，利用時精度が最も高いモデルが最も精度の高いモデルとして選択されるとは限らない，という結果になった．

6.2.　実験 2:提案手法による精度評価

提案手法による読み取り精度の評価結果を図 4 に示す．使用頻度が 50%の未満の範囲では，テスト対象の文字にはメジャーな文字よりもマイナーな文字のほうが多く含まれているため，実質的に意味を持つ評価結果は使用頻度が 50%よりも大きい範囲に限られる．また，図 3 における各モデルの読み取り精度と比較すると，全体的に低くなっている．

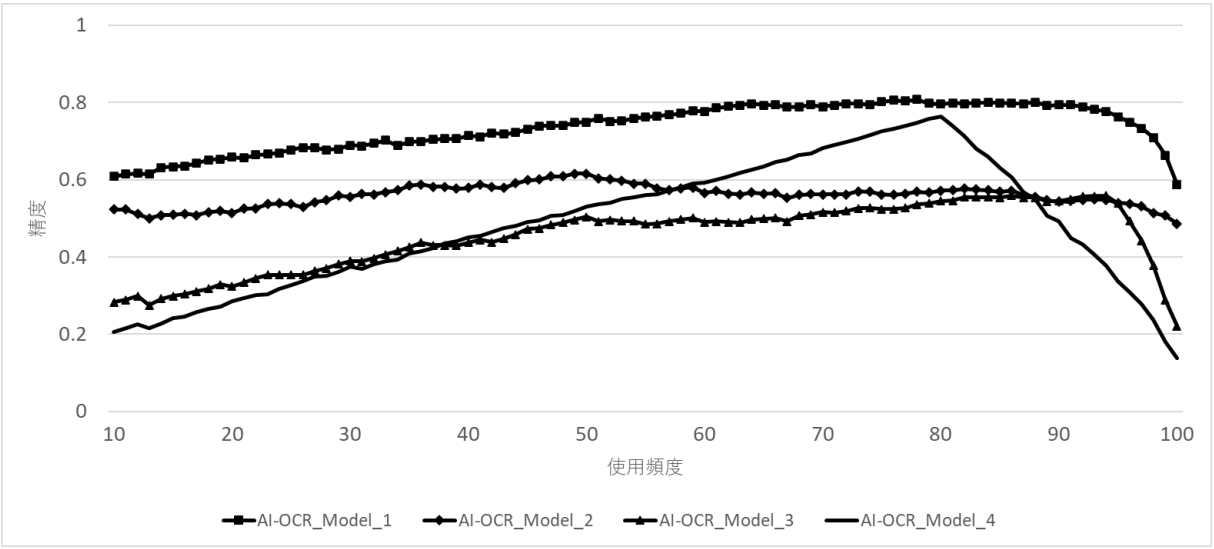

図 4　提案手法による読み取り精度の評価結果

AI-OCR_Model_4 における学習データを踏まえて，使用頻度の範囲を 50%から 80%までに限定して，提案手法による評価結果を確認すると，AI-OCR_Model_1 が最も精度が高いモデルであることがわかる．これは，4 つのモデルの中で利用時精度が最も高いことと，読み取り可能な文字の種類が最も多いことから,期待通りの結果である.また,AI-OCR_Model_2 と AI-OCR_Model_3 に着目すると，図 3 では利用時精度は AI-OCR_Model_3 のほうが AI-OCR_Model_2 よりも少し高いが,提案手法による評価結果では，AI-OCR_Model_2 のほうが AI-OCR_Model_3 よりも精度が高い．これは,AI-OCR_Model_2 のほうがより多くの文字に対応できていることが反映されていると考えられる.このように,提案手法による評価では，マイナーな文字を評価に含めることにより，利用時精度の順位が逆転することが確認できた．この結果についても，リスク回避性も含めた総合的な精度評価が行えていると考えられるため，期待通りの結果である．したがって，提案手法による評価では，AI パフォーマ

ンスとリスク回避性の両方をバランスよく評価できていると考えられる.

次に，提案手法によるテストデータの削減量について，テスト対象の全文字数との割合を図 5 に示す．提案手法におけるテストデータは，閾値までの範囲に含まれるすべてのメジャーな文字に，(4)式で示した数のマイナーな文字をランダムに付加している．したがって，図 5 では，マイナーな文字の削減量，つまり，テストデータとして選択されなかったマイナーな文字の数とすべての文字数の割合を示している．図 5 のグラフは，使用頻度が 80%程度までは削減割合は緩やかに減少し，それ以降では急激に減少していることがわかる．実験 2 で着目している使用頻度の範囲は 50%から 80%であり，削減量が緩やかに減少する範囲で評価を行っていることがわかる．また，この範囲における削減割合は 91.7%から 81.9%になっており，大幅なテストデータ削減を見込むことができる.

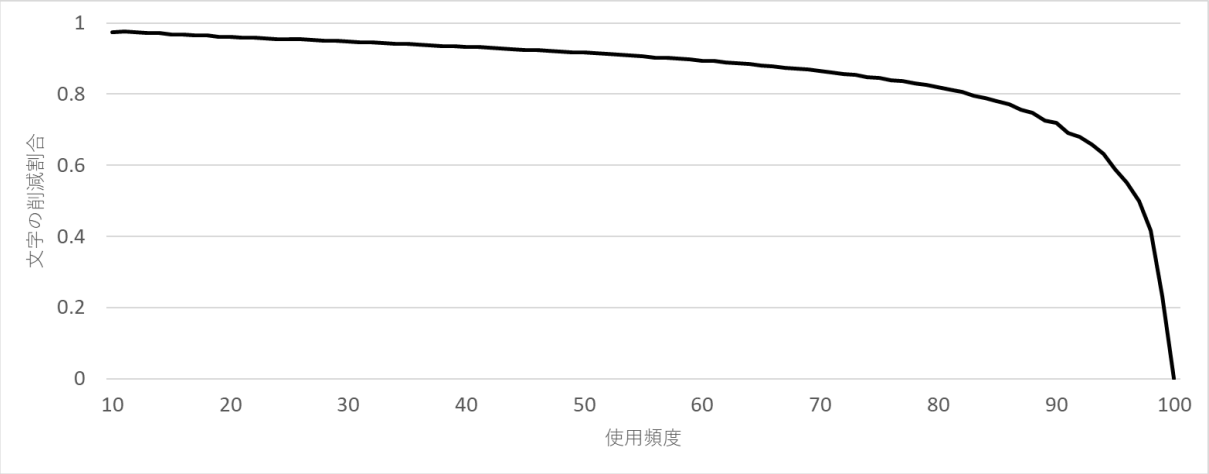

図 5　提案手法によるテスト対象文字の削減割合

6.3.　実験結果のまとめ

実験結果から，3 章で示した RQ に対する検証結果は以下の通りになる.

RQ.1-1　メジャーな文字のみを用いて読み取り精度を評価することで，利用時精度を予測することは可能である．ただし，予測が外れる場合も存在する

RQ.2-1　提案手法は，AI パフォーマンスとリスク回避性をバランスよく評価することが可能である

RQ.2-2 提案手法により，マイナーな文字に関してテストデータの大幅削減が可能である

7.　考察

7.1.　使用頻度が学習に与える影響

図 3 における読み取り精度の傾向として，使用頻度の高い文字ほど読み取り精度が高いことがわかる．これは，使用頻度の高い文字のほうが学習データの数が多いことが要因と考えられる．同じ理由で，マイナーな文字の読み取り精度は低くなる傾向にある．したがって，提案手法ではマイナーな文字を含めてテストしているため，実験 2 における各モデルの読み取り精度が実験 1 の読み取り精度よりも低下していると考えられる.

7.2.　メジャーな文字のみを評価することに対する考察

図 3 から，AI-OCR_Model_4 以外は 60%から 80%の範囲で利用時精度と読み取り精度の値が近いことがわかる．モデルごとに利用時精度と読み取り精度が完全一致するメジャーな文字の閾値は異なるが，この範囲であれば利用時精度を予測することが可能であると考えられる．また，閾値 80%における文字の割合は 14.5%なので，使用頻度の上位 2 割程度の文字を評価すれば，利用時精度を予測することが可能と考えられる.

一方，AI-OCR_Model_4 の利用時精度と読み取り精度の間に乖離が見られた原因として，学習していない文字が使用頻度の 20%という，無視できない程度に大きな割合であったことが挙げられる．また，学習している文字は 95%程度の精度と正解率が高く，学習してい

ない文字が全く読めないという正解率の二極化が起こっている．したがって，メジャーな文字のみを用いた評価では，正解率の高い領域のみを評価することになってしまった．その結果として，今回の乖離が生じたと考えることができる．

7.3.　提案手法による評価に対する考察

実験１と実験2における読み取り精度の差を図６に示す.ただし,今回は使用頻度が 50%以上の範囲のみを示す．図６から，実験２における読み取り精度は全体的に低いことが分かる．これは，提案手法ではマイナーな文字を付加していることと，マイナーな文字の読み取り精度はメジャーな文字の読み取り精度よりも低いことが要因と考えられる.したがって,マイナーな文字を付加することは，読み取れない文字に応じた，読み取り精度に対するペナルティになっている．

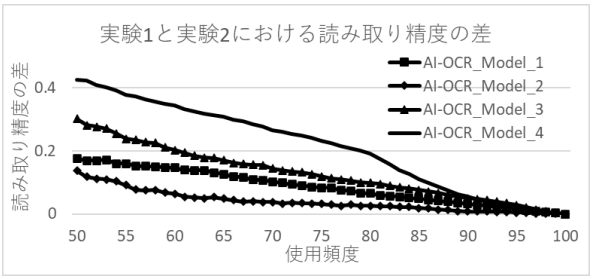

図６　実験１と実験２の読み取り精度の差

また，図 6 の 50%から 80%までの範囲では，AI-OCR_Model_4，AI-OCR_Model_3，AI-OCR_Model_1，AI-OCR_Model_2 の順で差が大きくなっている．特に前者の 2 つのモデルは一部の文字を学習していない．したがって，このグラフの結果からも，マイナーな文字を付加することで効果的なペナルティを科していることが確認できる．このペナルティの影響により，実験１と実験２における AI-OCR_Model_2 と AI-OCR_Model_3 の順位の入れ替わりが生じた．ただし，図３を確認すると，読み取り精度の差は僅差であるから，両モデルの読み取り精度の値が近いことも順位が逆転したことの一つの要因であると考えられる．

8.　今後の展望

本研究では，テスト対象に対して制限を設定して実験を行ったことも踏まえて，今後検証すべきこととして大きく２点が考えられる．

8.1. AI-OCR 以外の機械学習システムに対する提案手法の適用

今回扱った使用頻度の分布のように，インプットとなるデータがロングテイルな分布になっていることは，AI-OCR 以外にも広くみられる．AI-OCR 以外の分野で用いられる機械学習システムに対しても，提案手法が有効であることを検証する．また，AI-OCR 以外の様々な分野においても，本手法が適用できる有用な手法であることを検証する．

8.2. テスト対象の文字における制約の解除

今回は文字以外の要素を排除したが，実際の AI-OCR では，フォントや文字の大きさ，紙質，外乱などが読み取り精度に影響する．また，手書き文字についても，丁寧さや癖字などが読み取り精度に影響するため，これらの条件も含めて，提案手法による評価が有効であるかを検証する．

9.　謝辞

本研究を行うにあたり，多くの方々にお世話になりました．主査の石川冬樹氏，副主査の栗田太郎氏，徳本晋氏には，本研究を遂行する上で様々なご指導や助言をいただきました．深く感謝申し上げます．

参考文献

[1] 産業技術総合研究所，「機械学習品質マネジメントガイドライン 第３版」
[2] QA4AI コンソーシアム，「AI プロダクト品質保証ガイドライン，2022.07 版」
[3] 経済産業省，独立行政法人情報処理推進機構，「文字情報基盤 IPAmj」
[4] 文化庁，「漢字出現頻度数調査（4）（令和４年２月文化庁）の概要」
[5] 名字由来 net，https://myoji-yurai.net/
[6] https://github.com/tesseract-ocr/tesseract