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

第３８年度 研究コース５「人工知能とソフトウェア品質」（後藤銀行グループ）

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）

背景と課題

映像の振り返り

AI商社からの提案

モデルの名前	学習に使用したデータ	読み取り精度	順位
A社が作った万能型	秘伝のデータセットにつき非公開	0.861	2
日本の老舗ソフトウェアベンダーB社モデル	2245文字すべてを学習	0.598	4
イケイケ外資系C社モデル	2245文字のうち、名字でよく使用されている文字の95%である875文字	0.645	3
スタートアップのD社モデル	2245文字のうち、名字でよく使用されている文字の80%である325文字	**0.960**	**1**

果たして、D社のモデルを活用していいのだろうか？

後藤銀行のシステム担当者

背景と課題–出現頻度調査

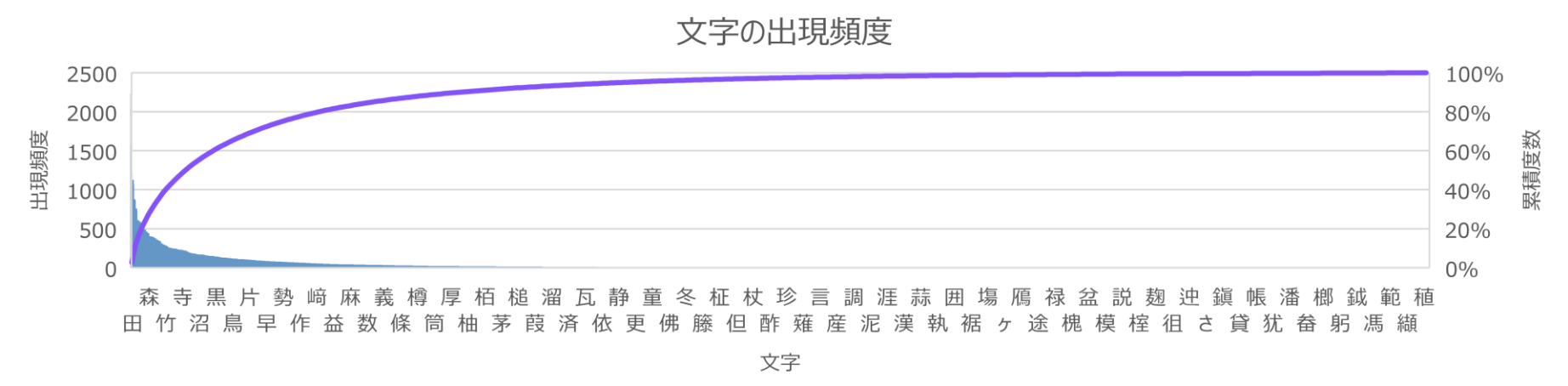
後藤王国国民全40,000人の名字の調査を行う

- 全国の名字における上位40,000件のデータを取得し、文字の出現頻度を集計

名字
中田
森田
:
田中
鈴木

出現頻度を集計

ID	文字	出現頻度
1	田	2230
2	森	1576
:	:	:
2244	鴛	1
2245	粧	1



出現頻度に偏りがあって、ロングテイルな分布になっていることが分かった

背景と課題-AI-OCRの精度について

精度の定義

- 読み取り精度：
  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。
  AI商社の営業が示した読み取り精度＝使用頻度80%までの文字の読み取り精度

- 利用時精度：
  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出現頻度で重みづけを行い、計算した割合

読み取り精度 p_t 、利用時精度 P は以下のように記述することができる

$$p_t = \frac{1}{n_t} \sum_{i \in \mathrm{char}_t} r_i$$

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

n_t：評価対象の文字数

char_t：評価対象の文字の集合

r_i：文字 i の読み取り結果

（正しく読み取れたら $r_i = 1$, それ以外は $r_i = 0$）

N：全文字数

$f(i)$：文字 i の使用頻度

背景と課題-読み取り精度とリスク回避性

AI商社から提案された4つのモデルに対して、提示された読み取り精度と追加で実施した評価結果

■ AI商社から提示された読み取り精度と利用時精度の比較

モデル	読み取り精度（80%）	利用時精度	差
A	0.861	0.859	+0.002
B	0.598	0.624	-0.026
C	0.645	0.652	-0.007
D	0.960	0.752	**+0.208**

読み取り精度と利用時精度に大きい差が生じることがある

■ 名字で使われる全2245文字を評価した結果

モデル	読み取り精度（100%）	リスク
A	0.587	中
B	0.486	中
C	0.221	**大**
D	0.138	**大**

読み取れない文字が多い＝リスク高

テスト設計に、読み取り精度とリスク回避性の考慮が必要である・・・課題①

背景と課題-AI-OCRのテストデータについて

評価対象の文字数が多いと、テストデータの作成にかかるコストが膨大になる

テストデータの数 ＝ 文字数 × フォントの数 × 文字の大きさ × 外乱 × …

日本語の場合
60,000字

- 漢字の出現頻度調査
  多数の文字は、ほとんど使われない文字である
- テストへの影響
  漢字の使用頻度が利用時精度へ影響する
- データ作成
  AI-OCRのテストデータは任意の文字・装飾・外乱等のデータに対して作成可能である

文字数を削減することでテストデータを効率よく作成することが必要・・・課題②

課題解決に向けたアプローチ

課題解決に向けたアプローチ-提案手法

出現頻度を考慮し、テストに使用する文字を取捨選択するテスト手法

特徴
ロングテイルなデータセットに対して、閾値ｔを決めて「メジャーな文字」「マイナーな文字」の2つのサブグループに分割する。
テストデータは、「メジャーな文字」は全ての文字を用いて、「マイナーな文字」の一部を削除することで作成する。
※ｔの範囲は、0≦ｔ≦100

考え方
・出現頻度の高い文字（メジャー）は、利用時精度への影響が大きいので、全て文字をテストに使用する。
・出現頻度の低い文字（マイナー）は、利用時精度への影響が小さいので、テストに使用する文字を削減する。

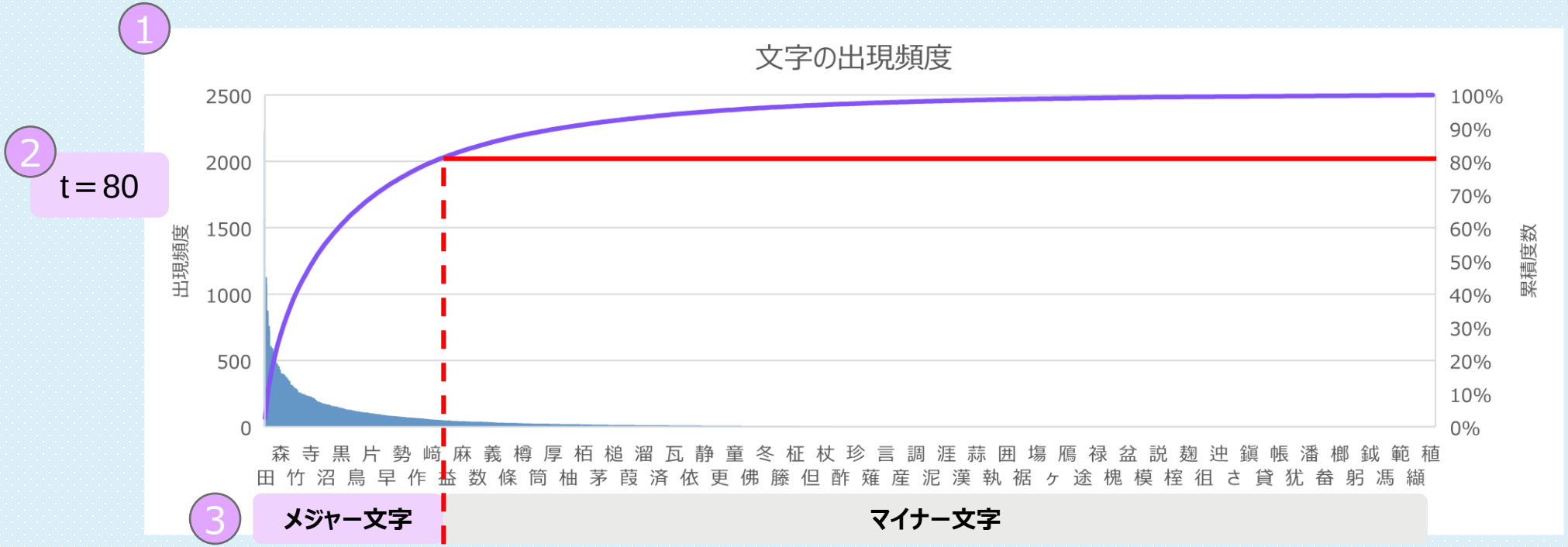
提案手法手順

1. ロングテイルなデータセットを用意する
2. 累積度数の閾値ｔを決める
3. 出現頻度上位t%をメジャー文字、出現頻度下位100-t%をマイナー文字としての2つのサブグループに分割する
4. メジャー文字：マイナー文字＝t：100-tになるようにマイナー文字を削除する

課題解決に向けたアプローチ-提案手法のフロー1/2

閾値 t を用いて、メジャー文字とマイナー文字を2つのサブグループに分割する

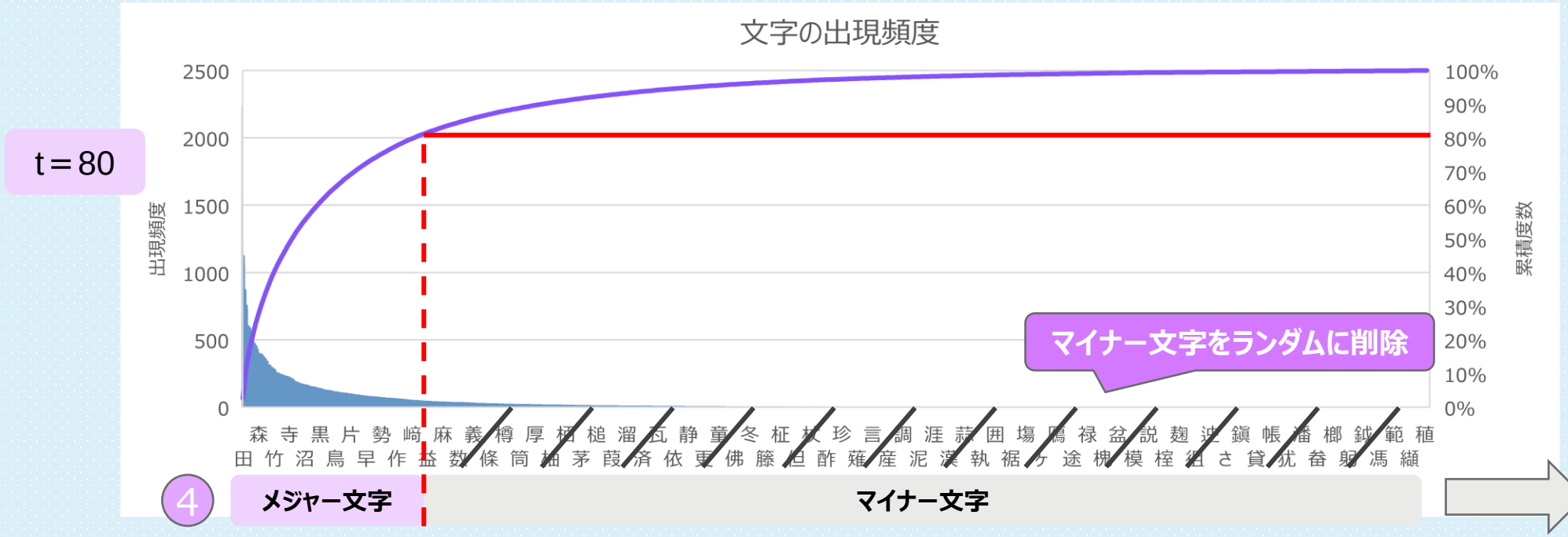
1. ロングテイルなデータセットを用意する
2. 累積度数の閾値 t を決める
3. 出現頻度上位t%をメジャー文字、出現頻度下位100-t%をマイナー文字としての2つのサブグループに分割する
4. メジャー文字：マイナー文字＝t：100-t になるようにマイナー文字を削除する



課題解決に向けたアプローチ-提案手法のフロー2/2

マイナーな文字をランダムに削減して、テストデータにおけるメジャー文字の数の比を閾値 t と一致させる

1. ロングテイルなデータセットを用意する
2. 累積度数の閾値 t を決める
3. 出現頻度上位t%をメジャー文字、出現頻度下位100-t%をマイナー文字としての2つのサブグループに分割する
4. メジャー文字：マイナー文字＝t：100-t になるようにマイナー文字を削除する



テストデータで使用する文字

メジャー文字	マイナー文字
80%	20%

実験

実験-提案手法の有効性の確認

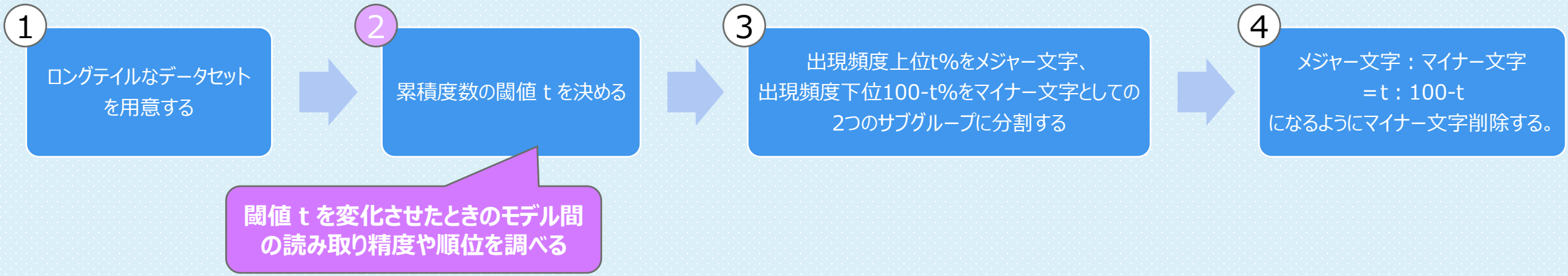
- **提案手法の有効性を確認するために、下記2点を検証する**

検証①：読み取り精度とリスク回避性のバランスが取れた評価が可能であること

1. 提案手法で評価したモデルの読み取り精度の順位を確認する。
   利用時精度とリスク回避性が最も優れているモデルが1位になることを確認する
2. 読み取り精度とリスク回避性を総合的に評価していること

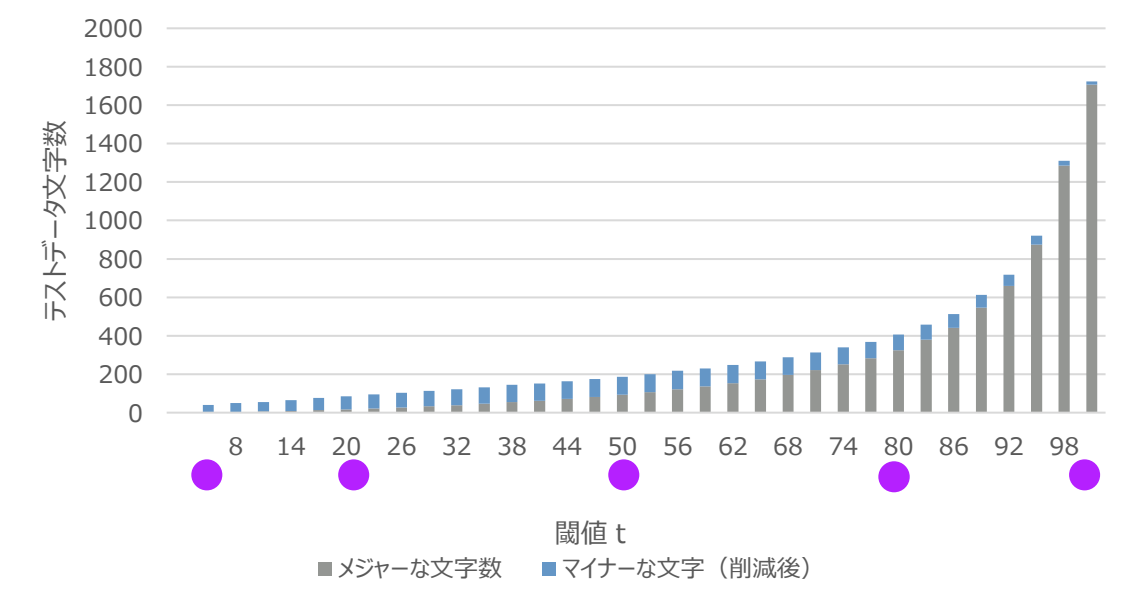
検証②：テストデータが大幅削減可能なこと

1. 全体の文字に対する削除したマイナーな文字が占める割合を確認する



実験前準備-閾値tの有効範囲

- 実験の前準備
  - 閾値 t とテストデータ数の確認



閾値 t	メジャー文字	マイナー文字（削除後）
0	0	0
20	17	68
50	93	93
80	325	81
100	2245	0

閾値 t について

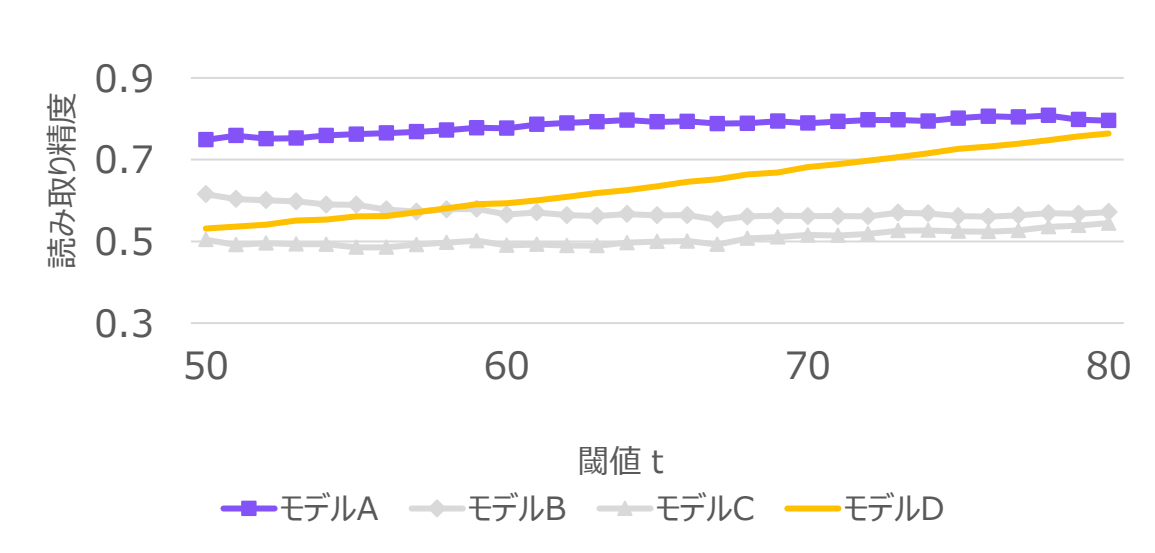
- 閾値50を境に、メジャー文字＞マイナー文字となる。
- 全てのモデルが学習している文字が含まれている範囲が80%までである。

評価が有効である範囲を「50 ≦ t ≦ 80」として確認していく

実験-検証①モデル間の読み取り精度について

■ 提案手法を用いた実験の結果

■ 閾値 t に対する読み取り精度



■ モデルごとの読み取り精度と利用時精度

モデル	読み取り精度 (80%)	利用時精度	読み取り精度 (100%)
A	0.796	0.859	0.587
B	0.571	0.624	0.486
C	0.545	0.652	0.221
D	0.764	0.752	0.138

分かったこと

・提案手法の有効範囲において、読み取り精度が「モデルA＞モデルD」であること
・利用時精度とリスク回避性ともに最も良いモデルAでは、提案手法を用いた読み取り精度（80%）も最も高くなった

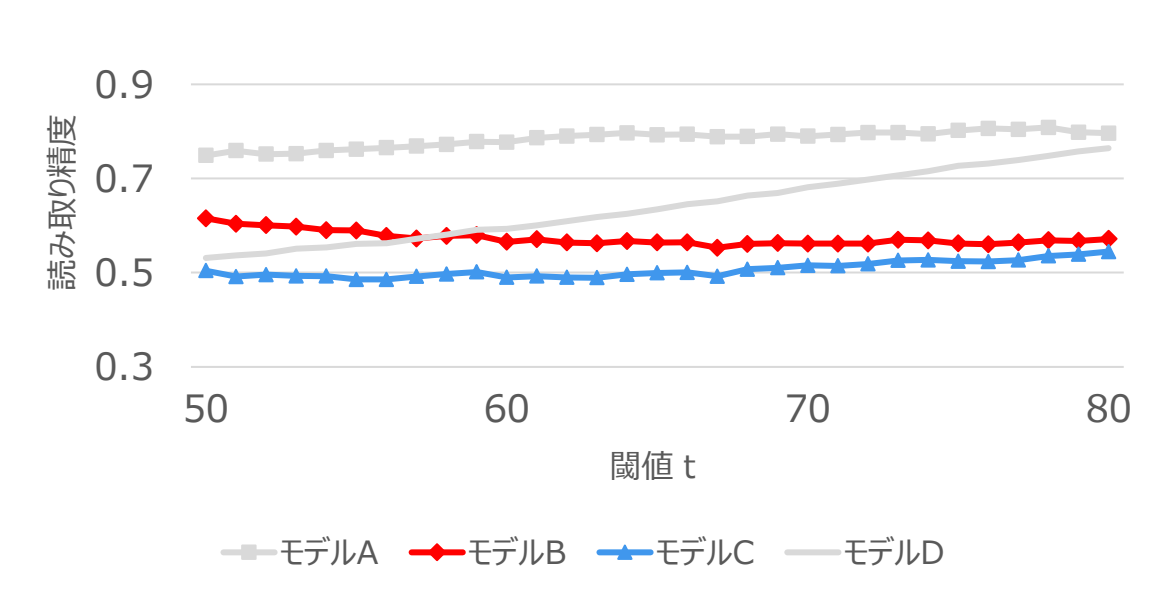
検証①-1「利用時精度とリスク回避性が最も優れているモデルが1位になることを確認する」への結果

・モデルAが最も良いモデルといえているので、
「利用時精度とリスク回避性が最も高いモデル」を「最も読み取り精度が高いモデル」として評価することができた

実験-検証①モデル間の読み取り精度について

■ 提案手法を用いた実験の結果

■ 閾値 t に対する読み取り精度



■ モデルごとの読み取り精度と利用時精度

モデル	読み取り精度 (80%)	利用時精度	読み取り精度 (100%)
A	0.796	0.859	0.587
B	0.571	0.624	0.486
C	0.545	0.652	0.221
D	0.764	0.752	0.138

分かったこと

・提案手法の有効範囲において、読み取り精度が「モデルB＞モデルC」であること

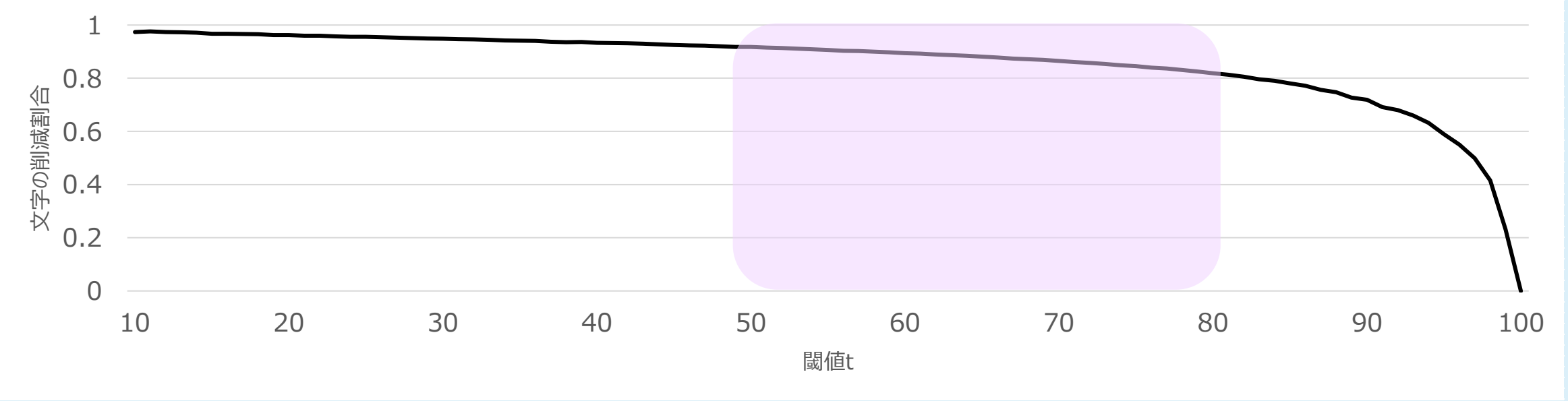
・利用時精度は、「モデルC＞モデルB」であり、リスク回避性は、「モデルB＞モデルC」であること

検証①－2「読み取り精度とリスク回避性を総合的に評価していること」

・モデルBとモデルCに対して、提案手法を用いることでリスク回避性を考慮することができた

実験-検証②文字の削減量について

- **提案手法を用いた実験の結果**
  - **テストデータの削減量**



分かったこと

・有効範囲である50%～80%の範囲では8割以上の文字削減を確認

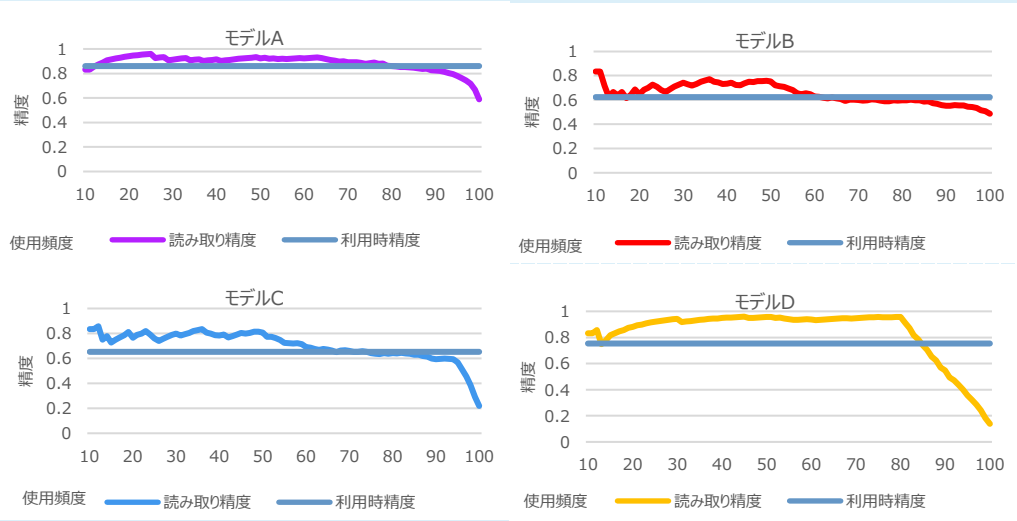
検証②－1「全体の文字に対する削除したマイナーな文字が占める割合を確認する」

・提案手法を用いるとテストデータの大幅削減が可能である

考察

考察

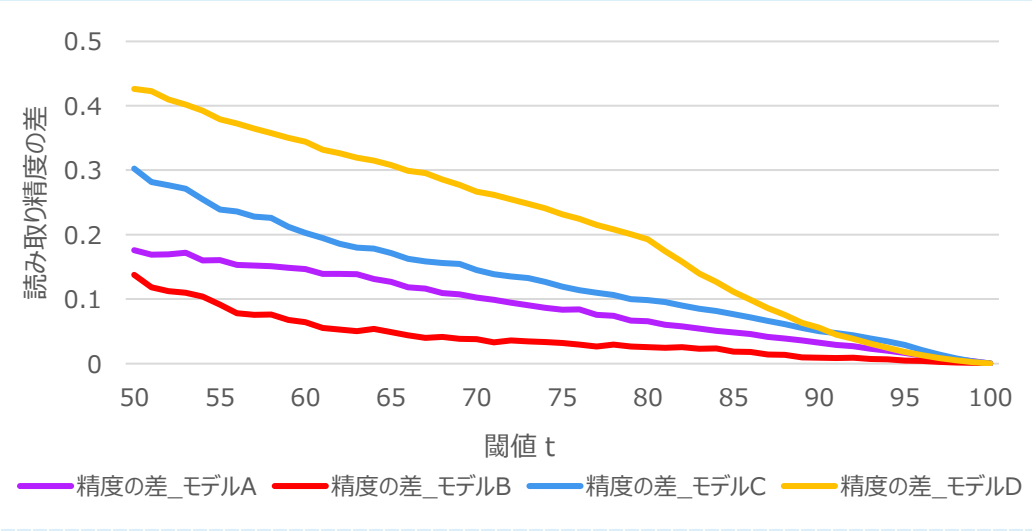
考察１：メジャーな文字のみを評価することに対する考察



モデルA、B、C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。一方，全体の14%の文字しか学習していないモデルDは大きな乖離が見られた。

メジャーな文字のみに特化させて学習させたモデルでは、読めないことに対する精度低下が無視できないほど大きい。

考察２：マイナーを付加することによる効果



1. 一般的にマイナーな文字のほうが読み取れない可能性が高い
2. マイナーな文字を付加することは読み取れない可能性の高い文字を評価に含めることになる
3. 読めない文字が多い（リスクが高い）ほどペナルティ(読み取り精度との差)が大きいので、リスク評価の役割を担っている

今後の展望

今後の展望

■ **今後の展望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**

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以下について検証する必要がある。

- AI-OCR 以外の分野で用いられる機械学習システムに対する有効性

例
- 総合ECサイトリコメンドシステム
  →「商品」と「売上」
- SEO対策
  →「言葉」と「アクセス数」

など

■ **今後の展望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。
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。

例
- 活字文字への外乱

- 手書き文字

D社モデルの性能は悪かったのだ
A社モデルが一番良かったのだ

ご清聴ありがとうなのだ！